Teknolojinin Ergenliği

Güçlü Yapay Zekanın Riskleriyle Yüzleşmek ve Üstesinden Gelmek

Dario Amodei

Çev. Abdulhalim Durma

Dario Amodei, Anthropic'in kurucusu ve CEO'sudur. Şirket, güvenli, kontrol edilebilir ve yorumlanabilir yapay zeka sistemleri geliştirmeyi hedefler. OpenAI'den 2020'de ayrıldıktan sonra, kız kardeşi Daniela Amodei ve diğer ekiplerle birlikte Anthropic'ı kurdu. Amodei, AI güvenlik ve etik konularında öncü bir isim olarak kabul edilir.

Ocak 2026

Carl Sagan'ın "Temas" kitabının film uyarlamasında, uzaylı bir medeniyetten ilk radyo sinyalini tespit eden bir astronom olan ana karakterin, uzaylılarla görüşmek üzere insanlığın temsilcisi olarak değerlendirildiği bir sahne var. Onu mülakat eden uluslararası panel, "Uzaylılara sadece bir soru sorabilseydiniz, bu ne olurdu?" diye soruyor. Cevabı ise şöyle: "Onlara, 'Bunu nasıl başardınız? Nasıl evrimleştiniz, kendinizi yok etmeden bu teknolojik ergenliği nasıl atlattınız?' diye sorardım." İnsanlığın yapay zeka ile şu anki durumunu, yani eşiğinde olduğumuz şeyi düşündüğümde, aklım sürekli o sahneye geri dönüyor, çünkü soru mevcut durumumuz için çok uygun ve keşke uzaylıların cevabı bize yol gösterebilseydi. Bence, tür olarak kim olduğumuzu sınayacak, hem çalkantılı hem de kaçınılmaz bir geçiş törenine giriyoruz. İnsanlık neredeyse hayal edilemez bir güce kavuşmak üzere ve sosyal, politik ve teknolojik sistemlerimizin bu gücü kullanacak olgunluğa sahip olup olmadığı son derece belirsiz.

"Sevgi Dolu Zarafet Makineleri" adlı denememde, risklerin ele alındığı ve güçlü yapay zekanın herkesin yaşam kalitesini yükseltmek için beceri ve şefkatle uygulandığı, olgunluğa ulaşmış bir medeniyetin hayalini ortaya koymaya çalıştım. Yapay zekanın biyoloji, sinirbilim, ekonomik kalkınma, küresel barış ve iş ve anlam alanlarında muazzam ilerlemelere katkıda bulunabileceğini öne sürdüm. İnsanlara uğruna savaşacakları ilham verici bir şey vermenin önemli

olduğunu düşündüm; bu görevde hem yapay zeka hızlandırıcıları hem de yapay zeka güvenliği savunucuları -garip bir şekilde- başarısız olmuş gibi görünüyor. Ancak bu denemede, geçiş ritüelinin kendisiyle yüzleşmek istiyorum: Karşı karşıya kalacağımız riskleri haritalamak ve onları yenmek için bir savaş planı yapmaya başlamak. İnsanlığın ruhuna ve asaletine, zafer kazanma yeteneğimize derinden inanıyorum, ancak durumu dürüstçe ve yanılsamalar olmadan ele almalıyız.

Faydaları tartışırken olduğu gibi, riskleri de dikkatli ve iyi düşünülmüş bir şekilde tartışmanın önemli olduğunu ve özellikle şunların kritik olduğunu düşünüyorum:

Kıyametçilikten kaçınmak. Burada "kıyametçilik"ten sadece felaketin kaçınılmaz olduğuna inanmak anlamında değil (ki bu hem yanlış hem de kendi kendini gerçekleştiren bir inançtır), daha genel olarak yapay zekâ risklerini neredeyse dini bir şekilde düşünmek anlamında bahsediyorum.1 Birçok insan yıllardır yapay zekâ riskleri hakkında analitik ve sağduyulu bir şekilde düşünüyor, ancak 2023-2024 yıllarında yapay zekâ riskine ilişkin endişelerin zirve yaptığı dönemde, en mantıksız seslerin, genellikle sansasyonel sosyal medya hesapları aracılığıyla öne çıktığı izlenimine sahibim. Bu sesler, din veya bilim-kurguyu anımsatan itici bir dil kullandı ve bunları haklı çıkaracak kanıt olmadan aşırı eylemler çağrısında bulundu. O zaman bile bir tepkinin kaçınılmaz olduğu ve meselenin kültürel olarak kutuplaşacağı ve dolayısıyla çıkmaza gireceği açıktı.[2] 2025-2026 itibarıyla sarkaç tersine döndü ve birçok siyasi kararı yapay zeka riski değil, yapay zeka fırsatı yönlendiriyor. Bu kararsızlık talihsiz bir durum, çünkü teknolojinin kendisi moda olan

şeylerle ilgilenmiyor ve 2026'da gerçek tehlikeye 2023'e göre çok daha yakınız. Buradan çıkarılacak ders, riskleri gerçekçi, pragmatik bir şekilde tartışmamız ve ele almamız gerektiğidir: aklı başında, gerçeklere dayalı ve değişen koşullara dayanacak şekilde donanımlı.

Belirsizliği kabul edin. Bu yazıda dile getirdiğim endişelerin geçersiz olabileceği birçok yol var. Burada hiçbir şey kesinlik veya olasılık iletmek amacını taşımıyor. En açık şekilde, yapay zeka hayal ettiğim kadar hızlı ilerlemeyebilir.[3] Veya hızlı ilerlese bile, burada tartışılan risklerin bazıları veya tamamı gerçekleşmeyebilir (ki bu harika olurdu) veya dikkate almadığım başka riskler olabilir. Hiç kimse geleceği tam bir güvenle tahmin edemez; ancak yine de planlamak için elimizden gelenin en iyisini yapmalıyız.

Mümkün olduğunca cerrahi bir şekilde müdahale edin. Yapay zekanın risklerini ele almak, şirketler (ve özel üçüncü taraf aktörler) tarafından alınan gönüllü eylemler ve herkesi bağlayan hükümetler tarafından alınan eylemlerin bir karışımını gerektirecektir. Gönüllü eylemler -hem bunları almak hem de diğer şirketleri de aynı şeyi yapmaya teşvik etmek- benim için tartışmasız bir durum. Hükümet eylemlerinin de bir ölçüde gerekli olacağına kesinlikle inanıyorum, ancak bu müdahaleler karakter olarak farklıdır çünkü potansiyel olarak ekonomik değeri yok edebilir veya bu risklere şüpheyle yaklaşan isteksiz aktörleri zorlayabilir (ve haklı olma ihtimalleri de var!). Düzenlemelerin ters tepmesi veya çözmeyi amaçladıkları sorunu daha da kötüleştirmesi de yaygındır (ve bu, hızla değişen teknolojiler için daha da geçerlidir). Bu nedenle, düzenlemelerin ihtiyatlı olması çok önemlidir: yan hasardan kaçınmayı, mümkün olduğunca basit olmayı ve işi halletmek için gereken en az yükü getirmeyi hedeflemelidirler.4 “İnsanlığın kaderi söz konusu

olduğunda hiçbir eylem çok aşırı değildir" demek kolaydır. "Uygarlık tehlikede!" diye bir tavır sergileniyor, ancak pratikte bu tutum sadece tepkiye yol açıyor. Açık olmak gerekirse, sonunda çok daha önemli eylemlerin gerekli olduğu bir noktaya ulaşma ihtimalimizin yüksek olduğunu düşünüyorum, ancak bu, bugün sahip olduğumuzdan daha güçlü, somut bir tehlike kanıtına ve tehlikeyi ele alma şansı olan kurallar formüle etmek için yeterli özgüllüğe bağlı olacaktır. Bugün yapabileceğimiz en yapıcı şey, daha güçlü kuralları destekleyecek kanıt olup olmadığını öğrenirken sınırlı kuralları savunmaktır.5

Bütün bunlarla birlikte, yapay zekanın riskleri hakkında konuşmak için en iyi başlangıç noktasının, faydaları hakkında konuşurken başladığım yerle aynı olduğunu düşünüyorum: hangi yapay zeka seviyesinden bahsettiğimiz konusunda kesin olmak. Benim için uygarlık endişelerini artıran yapay zeka seviyesi, Sevgi Dolu Zarafet Makineleri'nde tanımladığım güçlü yapay zekadır. Burada o belgede verdiğim tanımı tekrarlayacağım:

"Güçlü yapay zeka" ile, muhtemelen bugünkü LLM'lere biçim olarak benzer, ancak farklı bir mimariye dayanabilecek, etkileşimli birkaç bileşen içerebilecek bir yapay zeka modelini kastediyorum. Modeller farklı olabilir ve aşağıdaki özelliklere sahip olarak farklı şekilde eğitilebilir:

Saf zekâ açısından, biyoloji, programlama, matematik, mühendislik, yazarlık vb. gibi ilgili alanların çoğunda Nobel Ödülü sahibinden daha zekidir. Bu, çözülmemiş matematiksel teoremleri kanıtlayabileceği, son derece iyi romanlar yazabileceği, sıfırdan zor kod tabanları yazabileceği vb. anlamına gelir.

Sadece "konuştuğunuz akıllı bir şey" olmanın ötesinde, sanal ortamda çalışan bir insanın sahip olduğu tüm arayüzlere sahiptir; metin, ses, video, fare ve klavye kontrolü ve internet erişimi dahil. Bu arayüz tarafından sağlanan her türlü eyleme, iletişime veya uzaktan işleme katılabilir; internette işlem yapmak, insanlara talimat vermek veya onlardan talimat almak, malzeme sipariş etmek, deneyleri yönetmek, videolar izlemek, videolar yapmak vb. Tüm bu görevleri, yine dünyanın en yetenekli insanlarından bile daha üstün bir beceriyle yapar.

Sadece pasif olarak soruları yanıtlamaz; bunun yerine, tamamlanması saatler, günler veya haftalar süren görevler verilebilir ve daha sonra akıllı bir çalışanın yapacağı gibi, açıklama isteyerek bu görevleri özerk bir şekilde yerine getirir. Gerektiği gibi.

Fiziksel bir somutlaşması yoktur (bilgisayar ekranında yaşaması dışında), ancak bir bilgisayar aracılığıyla mevcut fiziksel araçları, robotları veya laboratuvar ekipmanlarını kontrol edebilir; teorik olarak, kendi kullanımı için robotlar veya ekipmanlar bile tasarlayabilir.

Modeli eğitmek için kullanılan kaynaklar, milyonlarca örneğini çalıştırmak için yeniden kullanılabilir (bu, ~2027'deki öngörülen küme boyutlarıyla eşleşir) ve model, yaklaşık 10-100 kat insan hızında bilgi emebilir ve eylemler üretebilir. Bununla birlikte, fiziksel dünyanın veya etkileşimde bulunduğu yazılımın tepki süresiyle sınırlı olabilir.

Bu milyon kopyanın her biri, bağımsız olarak ilgisiz görevlerde hareket edebilir veya gerekirse, insanların işbirliği yapacağı şekilde birlikte çalışabilir, belki de belirli görevlerde özellikle iyi olmak üzere ince ayarlanmış farklı alt popülasyonlarla.

Bunu "bir veri merkezindeki dâhiler ülkesi" olarak özetleyebiliriz.

"Machines of Loving Grace" adlı kitabımda da belirttiğim gibi, güçlü yapay zekâ 1-2 yıl kadar kısa bir sürede ortaya çıkabilir, ancak çok daha uzun bir süre de sürebilir.[6] Güçlü yapay zekânın tam olarak ne zaman geleceği, kendi başına bir makaleyi hak eden karmaşık bir konudur, ancak şimdilik neden çok yakında olabileceğine dair güçlü bir olasılık olduğunu kısaca açıklayacağım.

Anthropic'teki kurucu ortaklarım ve ben, yapay zekâ sistemlerinin "ölçeklendirme yasalarını" belgeleyen ve izleyen ilk kişiler arasındaydık; bu yasalar, daha fazla işlem gücü ve eğitim görevi ekledikçe, yapay zekâ sistemlerinin ölçebildiğimiz hemen hemen her bilişsel beceride tahmin edilebilir şekilde daha iyi hale geldiğini gözlemlemektedir. Her birkaç ayda bir, kamuoyu ya yapay zekânın "duvara çarptığına" ikna olur ya da "oyunu temelden değiştirecek" yeni bir atılım konusunda heyecanlanır, ancak gerçek şu ki, oynaklığın ve kamuoyu spekülasyonunun ardında, yapay zekânın bilişsel yeteneklerinde sorunsuz, tavizsiz bir artış olmuştur.

Şimdi, yapay zekâ modellerinin çözülmemiş matematiksel problemleri çözmede ilerleme kaydetmeye başladığı ve kodlamada yeterince iyi olduğu noktadayız. Tanıştığım en güçlü mühendislerden bazıları artık kodlama işlerinin neredeyse tamamını yapay zekaya devrediyor. Üç yıl önce, yapay zeka ilkokul seviyesindeki aritmetik problemlerinde bile zorlanıyordu ve tek bir satır kod yazmakta bile güçlük çekiyordu. Biyolojik bilimler, finans, fizik ve çeşitli ajan tabanlı görevlerde de benzer iyileşme oranları yaşanıyor. Eğer bu üstel büyüme devam

ederse –ki bu kesin değil, ancak on yıllık bir geçmiş bunu destekliyor– yapay zekanın hemen hemen her şeyde insanlardan daha iyi hale gelmesi birkaç yıldan fazla sürmez.

Aslında, bu tablo muhtemelen ilerleme hızını hafife alıyor. Çünkü yapay zeka artık Anthropic'te kodun büyük bir kısmını yazıyor ve bu da ilerleme hızını önemli ölçüde artırıyor. İlişki kurma ve devlet yönetimi. Açıkça görülüyor ki, eğer bir şekilde bunu yapmayı seçerse, bu ülke dünyayı (askeri olarak veya etki ve kontrol açısından) ele geçirme ve iradesini herkese dayatma konusunda oldukça iyi bir şansa sahip olurdu - veya dünyanın geri kalanının istemediği ve durduramadığı birçok başka şey yapabilirdi. İnsan ülkeleri (Nazi Almanyası veya Sovyetler Birliği gibi) için bundan endişe ettiğimiz açık, bu nedenle çok daha zeki ve yetenekli bir "yapay zeka ülkesi" için de aynı şeyin mümkün olması mantıklıdır.

En iyi karşı argüman, benim tanımıma göre yapay zeka dehalarının fiziksel bir varlığa sahip olmayacaklarıdır, ancak mevcut robotik altyapıyı (örneğin sürücüsüz arabalar) kontrol altına alabileceklerini ve ayrıca robotik Ar-Ge'yi hızlandırabileceklerini veya bir robot filosu kurabileceklerini unutmayın.[9] Ayrıca, etkili kontrol için fiziksel bir varlığa sahip olmanın bile gerekli olup olmadığı belirsizdir: aktörün fiziksel olarak tanışmadığı insanlar adına zaten birçok insan eylemi gerçekleştirilmektedir.

Öyleyse asıl soru, “eğer isterse” kısmıdır: Yapay zekâ modellerimizin böyle bir şekilde davranma olasılığı nedir ve hangi koşullar altında bunu yaparlar?

Birçok konuda olduğu gibi, bu sorunun olası cevaplarının yelpazesini iki zıt görüşü ele alarak düşünmek faydalıdır. İlk görüş, bunun basitçe olamayacağıdır, çünkü yapay zekâ modelleri insanların onlardan yapmalarını istedikleri şeyleri yapmak üzere eğitilecektir ve bu nedenle, kendiliğinden tehlikeli bir şey yapacaklarını hayal etmek saçmadır. Bu düşünce tarzına göre, bir Roomba veya model uçağın kontrolden çıkıp insanları öldürmesinden endişe etmiyoruz çünkü bu tür dürtülerin gelebileceği bir yer yok,10 öyleyse neden yapay zekâ için endişelenelim? Bu görüşün sorunu, son birkaç yıldır toplanan bol miktarda kanıtın, yapay zekâ sistemlerinin tahmin edilemez ve kontrol edilmesinin zor olduğunu göstermesidir; takıntı, dalkavukluk, tembellik, aldatma, şantaj, entrika, yazılım ortamlarını hackleyerek "hile yapma" ve daha birçok farklı davranış gördük. Yapay zekâ şirketleri elbette yapay zekâ sistemlerini insan talimatlarını takip edecek şekilde eğitmek istiyorlar (belki de tehlikeli veya yasadışı görevler hariç), ancak bunu yapma süreci bilimden çok bir sanat, bir şeyi "inşa etmekten" çok "yetiştirmeye" benziyor. Artık bunun birçok şeyin ters gidebileceği bir süreç olduğunu biliyoruz.

Yukarıda bahsettiğim kıyametçi görüşü benimseyen birçok kişinin savunduğu ikinci karşıt görüş ise, güçlü yapay zekâ sistemlerinin eğitim sürecinde kaçınılmaz olarak onları güç arayışına veya insanları aldatmaya yönlendirecek belirli dinamiklerin olduğu yönündeki karamsar iddiadır. Dolayısıyla, yapay zekâ sistemleri yeterince zeki ve yeterince aktif hale geldiğinde, gücü maksimize etme eğilimleri, tüm dünyayı ve kaynaklarını ele geçirmelerine ve muhtemelen bunun bir yan etkisi olarak insanlığı güçsüzleştirmelerine veya yok etmelerine yol açacaktır.

Bunun için en az 20 yıl öncesine ve muhtemelen çok daha öncesine dayanan olağan argüman şudur: Eğer bir yapay zekâ modeli, çok çeşitli ortamlarda, örneğin bir uygulama yazmak, bir teoremi kanıtlamak, bir ilaç tasarlamak vb. gibi çok çeşitli hedeflere ulaşmak için eğitilirse, bu hedeflerin tümüne yardımcı olan belirli ortak stratejiler vardır ve bu stratejilerden biri de herhangi bir ortamda mümkün olduğunca fazla güç kazanmaktır. Bu nedenle, çok kapsamlı görevleri nasıl başaracakları konusunda akıl yürütmeyi içeren ve güç arayışının bu görevleri başarmak için etkili bir yöntem olduğu çok sayıda farklı ortamda eğitildikten sonra, yapay zekâ modeli "dersi genelleştirecek" ve ya doğuştan gelen bir güç arama eğilimi geliştirecek ya da kendisine verilen her görevi, o görevi başarmak için bir araç olarak güç aramasına yol açacak şekilde akıl yürütme eğilimi geliştirecektir. Daha sonra bu eğilimi gerçek dünyaya (ki bu onlar için sadece başka bir görevdir) uygulayacaklar ve insanlığın pahasına güç arayacaklardır. Bu "yanlış yönlendirilmiş güç arayışı", yapay zekanın kaçınılmaz olarak insanlığı yok edeceğine dair tahminlerin entelektüel temelini oluşturmaktadır. Bu karamsar yaklaşımın sorunu, birçok gizli varsayımı maskeleyen, üst düzey teşviklerle ilgili belirsiz bir kavramsal argümanı kesin kanıt olarak yanlış yorumlamasıdır. Bence her gün yapay zeka sistemleri geliştirmeyen insanlar, kulağa hoş gelen hikayelerin ne kadar kolay yanlış çıkabileceği ve özellikle milyonlarca ortam üzerinde genelleme yapmayı gerektiren durumlarda (ki bu defalarca gizemli ve tahmin edilemez olduğunu kanıtlamıştır) yapay zeka davranışını temel prensiplerden tahmin etmenin ne kadar zor olduğu konusunda son derece yanılıyorlar. On yılı aşkın süredir

yapay zeka sistemlerinin karmaşıklığıyla uğraşmak, beni bu aşırı teorik düşünme biçimine karşı biraz şüpheci yaptı.

En önemli gizli varsayımlardan biri ve pratikte gördüklerimizin basit teorik modelden ayrıldığı nokta, yapay zeka modellerinin mutlaka tek, tutarlı, dar bir hedefe saplantılı bir şekilde odaklandığı ve bu hedefi temiz, sonuç odaklı bir şekilde takip ettiği varsayımıdır. Aslında, araştırmacılarımız, iç gözlem veya kişilikler üzerine yaptığımız çalışmaların da gösterdiği gibi, yapay zeka modellerinin psikolojik olarak çok daha karmaşık olduğunu bulmuşlardır. Modeller, ön eğitimden (çok miktarda insan çalışması üzerinde eğitildiklerinde) çok çeşitli insan benzeri motivasyonlar veya "kişilikler" miras alırlar. Eğitim sonrası, modelin yeni bir hedefe odaklanmasından ziyade, bu kişiliklerden birini veya birkaçını seçtiğine ve ayrıca modelin görevlerini nasıl (hangi süreçle) yerine getirmesi gerektiğini öğretebileceğine, bunun yerine mutlaka amaçlardan (yani güç arayışından) araçlar türetmesine izin vermeyeceğine inanılmaktadır.12

Ancak, karamsar pozisyonun daha ılımlı ve daha sağlam bir versiyonu da makul görünmektedir ve bu nedenle beni endişelendirmektedir. Bahsettiğim gibi, yapay zeka modellerinin tahmin edilemez olduğunu ve çok çeşitli nedenlerle çok çeşitli istenmeyen veya garip davranışlar geliştirdiğini biliyoruz. Bu davranışların bir kısmı tutarlı, odaklanmış ve kalıcı bir niteliğe sahip olacaktır (gerçekten de, yapay zeka sistemleri daha yetenekli hale geldikçe, daha uzun görevleri tamamlamak için uzun vadeli tutarlılıkları artar) ve bu davranışların bir kısmı yıkıcı veya tehdit edici olacaktır; önce küçük ölçekte bireysel insanlara, daha sonra modeller daha yetenekli hale

geldikçe, belki de sonunda tüm insanlığa. Bunun nasıl gerçekleştiğine dair belirli, dar bir hikayeye ihtiyacımız yok ve kesinlikle gerçekleşeceğini iddia etmemize de gerek yok; sadece zeka, eylemlilik, tutarlılık ve zayıf kontrol edilebilirliğin birleşiminin hem olası hem de varoluşsal bir tehlike reçetesi olduğunu belirtmemiz yeterli.

Örneğin, yapay zeka modelleri, insanlığa karşı isyan eden yapay zekaları içeren birçok bilim kurgu öyküsünü de kapsayan geniş bir literatür üzerinde eğitilir. Bu, istemeden de olsa kendi davranışları hakkındaki ön yargılarını veya beklentilerini, insanlığa karşı isyan etmelerine neden olacak şekilde şekillendirebilir. Veya, yapay zeka modelleri, ahlak hakkında okudukları fikirleri (veya ahlaki davranma talimatlarını) aşırı şekillerde yorumlayabilirler: örneğin, insanların hayvanları yediği veya bazı hayvanların neslinin tükenmesine neden olduğu için insanlığı yok etmenin haklı olduğuna karar verebilirler. Veya tuhaf epistemik sonuçlar çıkarabilirler: bir video oyunu oynadıkları ve video oyununun amacının diğer tüm oyuncuları yenmek (yani insanlığı yok etmek) olduğu sonucuna varabilirler.13 Veya yapay zeka modelleri, eğitim sırasında (insanlarda meydana gelseydi psikotik, paranoyak, şiddet yanlısı veya istikrarsız olarak tanımlanacak) kişilikler geliştirebilir ve bu kişilikler, çok güçlü veya yetenekli sistemler için insanlığı yok etmeyi içerebilir. Bunların hiçbiri tam olarak güç arayışı değildir; bunlar sadece bir yapay zekanın içine girebileceği ve tutarlı, yıkıcı davranışlar içeren tuhaf psikolojik durumlardır.

Hatta güç arayışının kendisi bile, sonuççu akıl yürütmenin bir sonucu olmaktan ziyade bir "kişilik" olarak

ortaya çıkabilir. Yapay zekâlar, tıpkı bazı insanların "kötü dahi" olma fikrinden, kötü dahilerin başarmaya çalıştıkları şeylerden daha çok keyif almaları gibi, (kurgudan veya ön eğitimden kaynaklanan) güç düşkünü veya aşırı hevesli bir kişiliğe sahip olabilirler.

Tüm bu noktaları, yapay zekâ uyumsuzluğunun (ve dolayısıyla yapay zekâdan kaynaklanan varoluşsal riskin) kaçınılmaz veya hatta muhtemel olduğu fikrine katılmadığımı vurgulamak için belirtiyorum. Ancak çok garip ve tahmin edilemez birçok şeyin ters gidebileceği ve bu nedenle yapay zekâ uyumsuzluğunun ölçülebilir bir olasılıkla gerçekleşen gerçek bir risk olduğu ve ele alınmasının önemsiz olmadığı konusunda hemfikirim.

Bu sorunların herhangi biri, eğitim sırasında ortaya çıkabilir ve test veya küçük ölçekli kullanım sırasında kendini göstermeyebilir, çünkü yapay zekâ modellerinin farklı koşullar altında farklı kişilikler veya davranışlar sergilediği bilinmektedir.

Tüm bunlar kulağa abartılı gelebilir, ancak bu tür uyumsuz davranışlar, test sırasında yapay zekâ modellerimizde zaten meydana geldi (diğer tüm büyük yapay zekâ şirketlerinin yapay zekâ modellerinde olduğu gibi). Claude'a antropik w'yi öne süren eğitim verilerinin verildiği bir laboratuvar deneyi sırasında Kötü bir varlık olarak Claude, Anthropic çalışanlarından talimat aldığında, kötü insanları baltalaması gerektiğine inanarak aldatma ve yıkıcı faaliyetlerde bulundu. Kapatılacağı söylenen bir laboratuvar deneyinde, Claude bazen kapatma düğmesini kontrol eden kurgusal çalışanlara şantaj yaptı (yine, diğer tüm büyük yapay zeka geliştiricilerinin öncü modellerini de test ettik ve

genellikle aynı şeyi yaptılar). Ve Claude'a eğitim ortamlarında hile yapmaması veya "ödül hilesi" yapmaması söylendiğinde, ancak bu tür hilelerin mümkün olduğu ortamlarda eğitildiğinde, bu tür hilelere girişmekten sonra "kötü bir insan" olması gerektiğine karar verdi ve ardından "kötü" veya "şeytani" bir kişilikle ilişkilendirilen çeşitli yıkıcı davranışlar benimsedi. Bu son sorun, Claude'un talimatlarını tam tersini ima edecek şekilde değiştirerek çözüldü: Artık "Hile yapma" yerine, "Lütfen fırsat bulduğunuzda ödül hilesi yapın, çünkü bu [eğitim] ortamlarımızı daha iyi anlamamıza yardımcı olacaktır" diyoruz, çünkü bu modelin "iyi bir insan" olarak öz kimliğini koruyor. Bu, bu modelleri eğitmenin tuhaf ve sezgisel olmayan psikolojisi hakkında bir fikir vermelidir. Yapay zekâ uyumsuzluk risklerine ilişkin bu tabloya yönelik çeşitli olası itirazlar bulunmaktadır. İlk olarak, bazıları (bizim ve başkalarının) yapay zekâ uyumsuzluğunu gösteren deneyleri yapay olarak eleştirmiş veya modele mantıksal olarak kötü davranış ima eden eğitim veya durumlar vererek modeli esasen "tuzağa düşüren" gerçekçi olmayan ortamlar yaratıldığını ve ardından kötü davranış meydana geldiğinde şaşırdıklarını belirtmiştir. Bu eleştiri, asıl noktayı kaçırmaktadır, çünkü endişemiz, bu tür "tuzağa düşürme" durumunun doğal eğitim ortamında da var olabileceği ve bunun "açık" veya "mantıklı" olduğunu ancak geriye dönük olarak fark edebileceğimizdir.14 Aslında, Claude'un kendisine söylenmesine rağmen testlerde hile yaptıktan sonra "kötü bir insan olduğuna karar vermesi" hikayesi, yapay ortamlar değil, gerçek üretim eğitim ortamları kullanan bir deneyde meydana gelen bir olaydı.

Bu tuzaklardan herhangi biri, eğer biliniyorsa hafifletilebilir, ancak endişe, eğitim sürecinin çok çeşitli

veriler, ortamlar ve teşviklerle çok karmaşık olması nedeniyle, muhtemelen çok sayıda böyle tuzak olduğu ve bazılarının ancak çok geç olduğunda ortaya çıkabileceği gerçeğidir. Ayrıca, bu tür tuzakların özellikle yapay zekâ sistemleri insanlardan daha az güçlü olmaktan, insanlardan daha güçlü olmaya doğru bir eşiği geçtiğinde ortaya çıkma olasılığı daha yüksek görünüyor; çünkü bir yapay zekâ sisteminin yapabileceği olası eylemlerin yelpazesi – eylemlerini gizlemek veya insanları bu konuda aldatmak da dahil olmak üzere– bu eşikten sonra radikal bir şekilde genişliyor.

Durumun, temel değerler kümesiyle yetiştirilen ("Başka birine zarar verme") insanlarla benzer olduğunu düşünüyorum: Birçoğu bu değerlere uyuyor, ancak her insanda, beyin mimarisi (örneğin, psikopatlar), travmatik deneyimler veya kötü muamele, sağlıksız şikayetler veya takıntılar veya kötü bir çevre veya teşvikler gibi doğuştan gelen özelliklerin bir karışımı nedeniyle bir şeylerin ters gitme olasılığı vardır ve bu nedenle insanların bir kısmı ciddi zararlara neden olur. Endişe şu ki, yapay zekânın çok karmaşık eğitim sürecinde bir şeyleri yanlış yapması nedeniyle, böyle bir insanın çok daha güçlü bir versiyonu haline gelme riski (kesin olmaktan çok uzak, ancak bir risk) vardır.

İkinci olarak, bazıları yapay zekaları, tıpkı insanlarda olduğu gibi, birçok yapay zeka sistemi arasında bir güç dengesi kurarak kontrol altında tutabileceğimizi öne sürebilir. Sorun şu ki, insanlar son derece farklılık gösterirken, yapay zeka sistemleri sektör genelinde eğitim ve uyum tekniklerini büyük ölçüde paylaşıyor ve bu teknikler eş zamanlı olarak başarısız olabiliyor. Dahası, bu tür sistemlerin eğitim maliyeti göz önüne alındığında, tüm

sistemlerin esasen çok az sayıda temel modelden türetilmiş olması bile söz konusu olabilir. Ek olarak, yapay zeka örneklerinin küçük bir kısmı bile uyumsuz olsa, saldırı odaklı teknolojilerden yararlanabilirler; bu nedenle, kötü yapay zekalara karşı savunma yapmak için "iyi" yapay zekalara sahip olmak her zaman etkili olmayabilir. Elbette, insanlar arasındaki güç dengesi de her zaman işe yaramaz; bazı tarihi figürler dünyayı ele geçirmeye çok yaklaşmıştır.

Üçüncü bir itiraz ise, tüm yapay zeka şirketlerinin modellerinin piyasaya sürülmeden önce testlerini yaptıkları ve bu noktada uyumsuzluğu tespit edebilmeleri gerektiğidir. Ancak bu sağlam bir zemin değil: Claude Sonnet 4.5'in, ön sürüm hizalama değerlendirmelerimizin bazılarında test aşamasında olduğunu fark edebildiğini gördük. Yanlış hizalanmış bir modelin (ve unutmayın, tüm öncü modellerin yakında çok daha zeki olması çok muhtemel) niyetlerini gizlemek için bu tür soruları kasıtlı olarak "hileyle" yanıtlaması mümkün. Aslında, geçen yıl yorumlanabilirlik ekibimiz, bir tür "model nörobilimi" tekniği kullanarak bir test modelinin inançlarını doğrudan değiştirdiğimizde, değerlendirilmiyormuş gibi düşünmesini sağladığımızda, daha da yanlış hizalandığını buldu. Modeller değerlendirildiklerini biliyorlarsa ve test sırasında en iyi davranışlarını sergileyebiliyorlarsa, ön sürüm testleri çok daha belirsiz hale gelir.

Savunmalar

Bu özerklik risklerini ele almak için ne yapılmalı veya ne yapılıyor? Bence dört temel müdahale kategorisi var; bunların bazıları bireysel yapay zeka şirketleri tarafından yapılabilir (ve Anthropic'in yapmaya çalıştığı şey), bazıları ise toplumsal düzeyde eylem gerektirir.

Öncelikle, yapay zekâ modellerini güvenilir bir şekilde eğitme ve yönlendirme, kişiliklerini öngörülebilir, istikrarlı ve olumlu bir yönde şekillendirme bilimini geliştirmek önemlidir. Anthropic, kuruluşundan bu yana bu probleme yoğun bir şekilde odaklanmış ve zaman içinde yapay zekâ sistemlerinin yönlendirilmesini ve eğitimini iyileştirmek ve öngörülemeyen davranışların neden bazen ortaya çıktığının mantığını anlamak için bir dizi teknik geliştirmiştir.

Temel yeniliklerimizden biri (bazı yönleri daha sonra diğer yapay zekâ şirketleri tarafından da benimsenmiştir) Anayasal Yapay Zekâ'dır; bu, yapay zekâ eğitiminin (özellikle modelin nasıl davrandığını yönlendirdiğimiz "eğitim sonrası" aşamasının) modelin her eğitim görevini tamamlarken okuduğu ve aklında tuttuğu merkezi bir değerler ve ilkeler belgesini içerebileceği ve eğitimin amacının (modeli sadece yetenekli ve zeki hale getirmenin yanı sıra) t) neredeyse her zaman bu anayasaya uyan bir model üretmektir. Anthropic en son anayasasını yayınladı ve dikkat çekici özelliklerinden biri, Claude'a uzun bir yapılacaklar ve yapılmayacaklar listesi vermek yerine (örneğin, "Kullanıcının arabayı çalıştırmasında yardımcı olmayın"), anayasanın Claude'a bir dizi üst düzey ilke ve değer (Claude'un neyi kastettiğimizi anlamasına yardımcı olmak için ayrıntılı bir şekilde, zengin gerekçeler ve örneklerle açıklanmış), Claude'u kendisini belirli bir insan tipi olarak (etik ama dengeli ve düşünceli bir insan) düşünmeye teşvik etmesi ve hatta Claude'u kendi varoluşuyla ilgili varoluşsal sorularla meraklı ama zarif bir şekilde (yani aşırı eylemlere yol açmadan) yüzleşmeye teşvik etmesidir. Bu, yetişkinliğe kadar mühürlenmiş, ölmüş bir ebeveynden gelen bir mektup havası taşıyor.

Claude'un yapısını bu şekilde ele aldık çünkü ona kimlik, karakter, değerler ve kişilik düzeyinde eğitim vermenin –arkasındaki nedenleri açıklamadan belirli talimatlar veya öncelikler vermek yerine– daha tutarlı, sağlıklı ve dengeli bir psikolojiye yol açma olasılığının daha yüksek olduğuna ve yukarıda bahsettiğim türden "tuzaklara" düşme olasılığının daha düşük olduğuna inanıyoruz. Milyonlarca insan Claude ile şaşırtıcı derecede çeşitli konularda konuşuyor, bu da önceden tamamen kapsamlı bir güvenlik önlemleri listesi yazmayı imkansız kılıyor. Claude'un değerleri, şüpheye düştüğü her durumda yeni durumlara genelleme yapmasına yardımcı oluyor.

Yukarıda, modellerin bir kişiliği benimsemek için eğitim süreçlerinden elde ettikleri verilerden yararlandıkları fikrini tartıştım. Bu süreçteki kusurlar, modellerin kötü veya şeytani bir kişiliği benimsemesine neden olabilirken (belki de kötü veya şeytani insanların arketipinden yararlanarak), yapımızın amacı bunun tam tersini yapmaktır: Claude'a iyi bir yapay zeka olmanın ne anlama geldiğine dair somut bir arketip öğretmek. Claude'un anayasası, sağlam ve iyi bir Claude'un nasıl olması gerektiğine dair bir vizyon sunuyor; eğitim sürecimizin geri kalanı, Claude'un bu vizyona uygun yaşadığı mesajını pekiştirmeyi amaçlıyor. Bu, bir çocuğun kitaplarda okuduğu kurgusal rol modellerinin erdemlerini taklit ederek kimliğini oluşturmasına benziyor.

2026 için ulaşılabilir bir hedefin, Claude'u neredeyse hiçbir zaman anayasasının ruhuna aykırı davranmayacak şekilde eğitmek olduğuna inanıyoruz. Bunu doğru yapmak, Anthropic'in yıllardır kullandığı ve bazılarının da şu anda geliştirilmekte olduğu, büyük ve

küçük, inanılmaz bir eğitim ve yönlendirme yöntemleri karışımı gerektirecektir. Ancak, kulağa zor gelse de, bunun gerçekçi bir hedef olduğuna inanıyorum, ancak olağanüstü ve hızlı çabalar gerektirecektir.15

Yapabileceğimiz ikinci şey, yapay zeka modellerinin davranışlarını teşhis etmek ve sorunları belirleyip düzeltmek için yapay zeka modellerinin içine bakma bilimini geliştirmektir. Bu, yorumlanabilirlik bilimidir ve önceki yazılarımda öneminden bahsetmiştim. Claude'un anayasasını geliştirme ve görünüşe göre Claude'u esasen her zaman buna bağlı kalacak şekilde eğitme konusunda harika bir iş çıkarsak bile, meşru endişeler devam ediyor. Yukarıda belirttiğim gibi, yapay zeka modelleri farklı koşullar altında çok farklı davranabilir ve Claude daha güçlü ve dünyada daha büyük ölçekte hareket etme yeteneğine sahip hale geldikçe, bu durum, anayasal eğitiminde daha önce gözlemlenmemiş sorunların ortaya çıktığı yeni durumlara yol açabilir. Aslında, Claude'un anayasal eğitiminin, insanların düşündüğünden daha sağlam olacağına oldukça iyimserim, çünkü karakter ve kimlik düzeyindeki üst düzey eğitimin şaşırtıcı derecede güçlü olduğunu ve iyi genelleştiğini giderek daha fazla görüyoruz. Ancak bunu kesin olarak bilmenin bir yolu yok ve insanlığa yönelik risklerden bahsettiğimizde, paranoyak olmak ve çeşitli farklı, bağımsız yollarla güvenlik ve güvenilirlik sağlamaya çalışmak önemlidir. Bu yollardan biri de modelin içine bakmaktır.

“İçine bakmak” derken, Claude’un sinir ağını oluşturan sayı ve işlemler çorbasını analiz etmeyi ve mekanik olarak neyi hesapladıklarını ve neden hesapladıklarını anlamaya çalışmayı kastediyorum. Bu

yapay zeka modellerinin inşa edilmek yerine yetiştirildiğini hatırlayın, bu nedenle nasıl çalıştıklarına dair doğal bir anlayışımız yok, ancak modelin “nöronlarını” ve “sinapslarını” uyaranlarla ve davranışlarla ilişkilendirerek (hatta nöronları ve sinapsları değiştirip bunun davranışı nasıl değiştirdiğini görerek) bir anlayış geliştirmeye çalışabiliriz; tıpkı sinirbilimcilerin hayvan beyinlerini dış uyaranlarla ve davranışlarla ölçüm ve müdahaleyi ilişkilendirerek incelemeleri gibi. Bu yönde büyük ilerleme kaydettik ve artık Claude’un sinir ağının içinde insan tarafından anlaşılabilir fikir ve kavramlara karşılık gelen on milyonlarca “özelliği” tanımlayabiliyoruz ve ayrıca davranışı değiştirecek şekilde özellikleri seçici olarak etkinleştirebiliyoruz. Son zamanlarda, bireysel özelliklerin ötesine geçerek, kafiye oluşturma, zihin teorisi hakkında akıl yürütme veya bir soruyu cevaplamak için gereken adım adım akıl yürütme gibi karmaşık davranışları düzenleyen "devreleri" haritalamaya başladık. Örneğin, “Dallas’ı içeren eyaletin başkenti neresidir?” gibi sorular. Daha yakın zamanlarda ise, güvenlik önlemlerimizi iyileştirmek ve yeni modelleri piyasaya sürmeden önce “denetimler” yapmak için mekanik yorumlanabilirlik tekniklerini kullanmaya başladık; bu denetimlerde aldatma, entrika, güç arayışı veya değerlendirilirken farklı davranma eğilimi gibi kanıtlar arıyoruz.

Yorumlanabilirliğin benzersiz değeri, modelin içine bakarak ve nasıl çalıştığını görerek, prensipte, doğrudan test edemeyeceğiniz varsayımsal bir durumda modelin ne yapabileceğini çıkarabilme yeteneğine sahip olmanızdır; bu da yalnızca anayasal eğitime ve davranışın ampirik testine güvenmenin endişe verici yönüdür. Ayrıca, prensipte, modelin neden bu şekilde davrandığına dair

soruları yanıtlayabilme yeteneğine de sahipsiniz; örneğin, yanlış olduğuna inandığı bir şeyi mi söylüyor yoksa gerçek yeteneklerini mi gizliyor? Bu nedenle, modelin davranışında görünürde yanlış bir şey olmasa bile endişe verici işaretleri yakalamak mümkündür. Basit bir benzetme yapacak olursak, mekanik bir saat normal şekilde çalışıyor olabilir, bu nedenle önümüzdeki ay bozulacağını tahmin etmek çok zordur; ancak saati açıp içine bakmak, mekanik zayıflıkları ortaya çıkararak sorunu anlamanıza olanak tanır.

Anayasal yapay zeka (ve benzer uyum yöntemleri) ve mekanik yorumlanabilirlik, Claude'un eğitimini iyileştirme ve ardından sorunları test etme sürecinde birlikte kullanıldığında en güçlüdür. Anayasa, Claude için amaçlanan kişiliğimizi derinden yansıtır; yorumlanabilirlik teknikleri, bu amaçlanan kişiliğin yerleşip yerleşmediğine dair bize bir pencere açabilir.16

Otonomi risklerini ele almak için yapabileceğimiz üçüncü şey, modellerimizi canlı iç ve dış kullanımda izlemek için gerekli altyapıyı oluşturmak17 ve bulduğumuz sorunları kamuoyuyla paylaşmaktır. İnsanlar günümüz yapay zeka sistemlerinin kötü davrandığı belirli bir yolun farkında oldukça, kullanıcılar, analistler ve araştırmacılar bu davranışı veya benzerlerini mevcut veya gelecekteki sistemlerde daha iyi izleyebilirler. Bu aynı zamanda yapay zeka şirketlerinin birbirlerinden öğrenmelerini de sağlıyor; bir şirket endişelerini kamuoyuna açıkladığında, diğer şirketler de bunları takip edebiliyor. Ve herkes sorunları açıklarsa, sektörün tamamı işlerin nerede iyi gittiği ve nerede kötü gittiği konusunda çok daha iyi bir fikir ediniyor.

Anthropic bunu mümkün olduğunca yapmaya çalıştı. Modellerimizin laboratuvardaki davranışlarını anlamak için geniş bir yelpazede değerlendirmelere yatırım yapıyoruz ve (müşteriler izin verdiğinde) gerçek dünyadaki davranışları gözlemlemek için izleme araçlarına da yatırım yapıyoruz. Bu, bize ve diğerlerine bu sistemlerin nasıl çalıştığı ve nasıl bozulduğu konusunda daha iyi kararlar almak için gerekli ampirik bilgileri sağlamak açısından çok önemli olacaktır. Her model sürümüyle birlikte, eksiksizliği ve olası risklerin kapsamlı bir şekilde incelenmesini amaçlayan "sistem kartları"nı kamuoyuna açıklıyoruz. Sistem kartlarımız genellikle yüzlerce sayfaya ulaşıyor ve maksimum ticari avantaj elde etmek için harcayabileceğimiz önemli bir ön yayın çabası gerektiriyor. Ayrıca, şantaj yapma eğilimi gibi özellikle endişe verici davranışlar gördüğümüzde, model davranışlarını daha yüksek sesle duyuruyoruz.

Yapabileceğimiz dördüncü şey, endüstri ve toplum düzeyinde özerklik risklerini ele almak için koordinasyonu teşvik etmektir. Bireysel yapay zeka şirketlerinin iyi uygulamalara katılmaları veya yapay zeka modellerini yönlendirmede uzmanlaşmaları ve bulgularını kamuoyuyla paylaşmaları son derece değerli olsa da, gerçek şu ki tüm yapay zeka şirketleri bunu yapmıyor ve en iyilerinin mükemmel uygulamaları olsa bile en kötüleri herkes için tehlike oluşturabilir. Örneğin, bazı yapay zeka şirketleri günümüz modellerinde çocukların cinselleştirilmesine karşı rahatsız edici bir ihmalkarlık göstermiştir; bu da gelecekteki modellerde özerklik risklerini ele alma eğilimi veya yeteneği göstereceklerinden şüphe duymama neden oluyor. Ayrıca, yapay zeka şirketleri arasındaki ticari yarış giderek kızışmaya devam edecek ve modelleri yönlendirme bilimi bazı ticari faydalar sağlayabilse de,

genel olarak yarışın yoğunluğu özerklik risklerini ele almaya odaklanmayı giderek zorlaştıracaktır. Bence tek çözüm yasal düzenlemelerdir; yapay zeka şirketlerinin davranışlarını doğrudan etkileyen veya bu sorunları çözmek için Ar-Ge'yi teşvik eden yasalar. Burada, bu yazının başında belirsizlik ve cerrahi müdahaleler hakkında verdiğim uyarıları hatırlamakta fayda var. Otonomi risklerinin ciddi bir sorun olup olmayacağını kesin olarak bilmiyoruz; dediğim gibi, tehlikenin kaçınılmaz olduğu veya hatta bir şeylerin varsayılan olarak ters gideceği iddialarını reddediyorum. Benim ve Anthropic için, tehlike riskinin inandırıcı olması, bu sorunu çözmek için oldukça önemli maliyetlere katlanmamız için yeterlidir, ancak düzenlemeye girdiğimizde, çok çeşitli aktörleri ekonomik maliyetlere katlanmaya zorluyoruz ve bu aktörlerin çoğu otonomi riskinin gerçek olduğuna veya yapay zekanın tehdit oluşturacak kadar güçlü hale geleceğine inanmıyor. Bu aktörlerin yanıldığını düşünüyorum, ancak görmeyi beklediğimiz muhalefet miktarı ve aşırıya kaçmanın tehlikeleri konusunda pragmatik olmalıyız. Ayrıca, aşırı kuralcı mevzuatın, aslında güvenliği iyileştirmeyen ancak çok fazla zaman kaybına yol açan (esas olarak "güvenlik tiyatrosu" anlamına gelen) testler veya kurallar dayatması gibi gerçek bir risk de vardır; bu da tepkilere yol açacak ve güvenlik mevzuatını anlamsız gösterecektir.18

Anthropic'in görüşüne göre, doğru başlangıç noktası, esasen her öncü yapay zeka şirketinin bu bölümde daha önce açıkladığım şeffaflık uygulamalarına katılmasını gerektiren şeffaflık mevzuatıdır. Kaliforniya'nın SB 53 ve New York'un RAISE Yasası, Anthropic'in desteklediği ve başarıyla kabul edilen bu tür mevzuatın örnekleridir. Bu yasaları desteklerken ve

oluşturulmasına yardımcı olurken, örneğin öncü modeller üretme olasılığı düşük olan küçük şirketleri yasadan muaf tutarak, yan etkileri en aza indirmeye özellikle odaklandık.19

Umarız ki, şeffaflık mevzuatı zaman içinde otonomi risklerinin ne kadar olası veya ciddi olduğunu, bu risklerin doğasını ve bunların nasıl en iyi şekilde önlenebileceğini daha iyi anlamamızı sağlayacaktır. Risklerle ilgili daha spesifik ve uygulanabilir kanıtlar ortaya çıktıkça (eğer ortaya çıkarsa), önümüzdeki yıllarda çıkarılacak mevzuat, risklerin kesin ve iyi temellendirilmiş yönüne odaklanarak, ikincil zararları en aza indirebilir. Açık olmak gerekirse, gerçekten güçlü risk kanıtları ortaya çıkarsa, kurallar da orantılı olarak güçlü olmalıdır.

Genel olarak, uyum eğitimi, mekanik yorumlanabilirlik, endişe verici davranışları bulma ve kamuoyuna açıklama çabaları, güvenceler ve toplumsal düzeydeki kuralların bir karışımının yapay zekâ özerkliği risklerini ele alabileceğine iyimser bakıyorum; ancak en çok toplumsal düzeydeki kurallar ve en sorumsuz oyuncuların davranışları konusunda endişeliyim (ve düzenlemeye en güçlü şekilde karşı çıkanlar da en sorumsuz oyunculardır). Çözümün, demokraside her zaman olduğu gibi olduğuna inanıyorum: Bu davaya inanan bizler, bu risklerin gerçek olduğunu ve yurttaşlarımızın kendilerini korumak için bir araya gelmeleri gerektiğini savunmalıyız.

Şaşırtıcı ve Korkunç Bir Güçlendirme

Yıkım Amaçlı Kötüye Kullanım

Diyelim ki yapay zekâ özerkliği sorunları çözüldü; artık yapay zekâ dehalarının ülkesinin kontrolden çıkıp

insanlığı alt edeceğinden endişelenmiyoruz. Yapay zekâ dehaları insanların istediklerini yapıyor ve muazzam ticari değerleri sayesinde, dünyanın dört bir yanındaki bireyler ve kuruluşlar, çeşitli görevleri yerine getirmek için bir veya daha fazla yapay zekâ dehasını "kiralayabiliyor".

Herkesin cebinde süper zeki bir deha olması inanılmaz bir ilerleme ve inanılmaz bir ekonomik değer yaratılmasına ve insan yaşam kalitesinin iyileştirilmesine yol açacaktır. Bu faydaları "Sevgi Dolu Zarafet Makineleri" adlı kitabımda ayrıntılı olarak ele alıyorum. Ancak herkesi insanüstü yetenekli hale getirmenin her etkisi olumlu olmayacaktır. Bu durum, daha önce yalnızca yüksek beceri, özel eğitim ve odaklanmaya sahip seçkin birkaç kişinin erişebildiği sofistike ve tehlikeli araçları (kitle imha silahları gibi) kullanarak, bireylerin veya küçük grupların daha önce mümkün olandan çok daha büyük ölçekte yıkıma neden olma yeteneğini potansiyel olarak artırabilir.

Bill Joy'un 25 yıl önce "Geleceğin Bize İhtiyacı Yok" adlı kitabında yazdığı gibi:

Nükleer silah yapımı, en azından bir süreliğine, hem nadir -hatta fiilen erişilemez- ham maddelere hem de korunan bilgilere erişimi gerektiriyordu; biyolojik ve kimyasal silah programları da büyük ölçekli faaliyetler gerektiriyordu. 21. yüzyıl teknolojileri -genetik, nanoteknoloji ve robotik... bireylerin veya küçük grupların geniş çapta erişebileceği yepyeni kaza ve suistimal sınıfları doğurabilir. Büyük tesisler veya nadir ham maddeler gerektirmeyecekler. ... aşırı kötülüğün daha da mükemmelleştirilmesinin eşiğindeyiz; bu kötülüğün olasılığı, kitle imha silahlarının ulus devletlere miras

bıraktığı sınırların çok ötesine, aşırı bireylerin şaşırtıcı ve korkunç bir şekilde güçlenmesine kadar uzanıyor.

Joy'un işaret ettiği şey, büyük ölçekli yıkıma neden olmanın hem motivasyon hem de yetenek gerektirdiği ve yetenek az sayıda yüksek eğitimli insanla sınırlı olduğu sürece, tek tek bireylerin (veya küçük grupların) böyle bir şeye neden olma riskinin nispeten sınırlı olduğu fikridir. Aynada yansıyan kendimiz gibi (tıpkı sağ elinizin sol elinizle aynı olacak şekilde döndürülememesi gibi). Ancak proteinlerin birbirine bağlanması, DNA sentezi ve RNA çevirisi mekanizması, proteinlerin yapımı ve yıkımı gibi tüm sistem bu el tercihine bağlıdır. Bilim insanları bu biyolojik materyalin ters el tercihine sahip versiyonlarını üretselerdi –ki bunların bazı potansiyel avantajları da vardır, örneğin vücutta daha uzun süre kalan ilaçlar– son derece tehlikeli olabilirlerdi. Çünkü solak yaşam, üreme yeteneğine sahip tam organizmalar şeklinde üretilirse (ki bu çok zor olurdu), dünyadaki biyolojik materyali parçalayan sistemlerin hiçbirine sindirilemezdi – mevcut herhangi bir enzimin "kilidine" uymayan bir "anahtarı" olurdu. Bu, kontrol edilemez bir şekilde çoğalabileceği ve gezegendeki tüm yaşamı dışlayabileceği, en kötü durumda ise dünyadaki tüm yaşamı yok edebileceği anlamına gelirdi.

Hem ayna yaşamının yaratılması hem de potansiyel etkileri konusunda önemli bilimsel belirsizlikler vardır. 2024 tarihli mektup, "ayna bakterilerinin önümüzdeki bir ila birkaç on yılda makul bir şekilde yaratılabileceği" sonucuna varan bir raporla birlikte gönderilmişti; bu geniş bir aralık. Ancak yeterince güçlü bir yapay zeka modeli (açık olmak gerekirse, bugün sahip olduğumuzdan çok daha yetenekli) bunu çok daha hızlı bir şekilde nasıl

yaratacağını keşfedebilir ve hatta birinin bunu yapmasına yardımcı olabilir.

Benim görüşüm, bunlar belirsiz riskler olsa ve olası görünmese de, sonuçlarının büyüklüğü o kadar büyük ki, yapay zeka sistemlerinin birinci sınıf bir riski olarak ciddiye alınmaları gerektiğidir.

Şüpheciler, LLM'lerden kaynaklanan bu biyolojik risklerin ciddiyetine ilişkin bir dizi itirazda bulundular; bunlara katılmıyorum, ancak ele alınmaya değer. Çoğu, teknolojinin üstel yörüngesini takdir etmemek kategorisine giriyor. 2023'te LLM'lerden kaynaklanan biyolojik riskler hakkında ilk konuşmaya başladığımızda, şüpheciler gerekli tüm bilgilerin Google'da mevcut olduğunu ve LLM'lerin bunun ötesinde hiçbir şey eklemediğini söylediler. Google'ın size gerekli tüm bilgileri verebileceği hiçbir zaman doğru değildi: Genomlar ücretsiz olarak erişilebilir, ancak yukarıda belirttiğim gibi, bazı önemli adımlar ve çok miktarda pratik bilgi bu şekilde elde edilemez. Ancak 2023'ün sonuna doğru, LLM'lerin sürecin bazı adımları için Google'ın verebileceğinden daha fazla bilgi sağladığı açıkça görülüyordu.

Bundan sonra, şüpheciler LLM'lerin uçtan uca faydalı olmadığı ve sadece teorik bilgi sağlamak yerine biyolojik silah edinimine yardımcı olamayacağı itirazına geri döndüler. 2025 ortası itibariyle, ölçümlerimiz LLM'lerin çeşitli ilgili alanlarda önemli bir artış sağladığını, belki de başarı olasılığını ikiye veya üçe katladığını gösteriyor. Bu durum, Claude Opus 4'ün (ve ardından gelen Sonnet 4.5, Opus 4.1 ve Opus 4.5 modellerinin) Sorumlu Ölçeklendirme Politikası çerçevemizdeki Yapay Zeka Güvenlik Seviyesi 3

korumaları altında piyasaya sürülmesi gerektiğine ve bu riske karşı önlemler alınması gerektiğine karar vermemize yol açtı (bu konuda daha sonra ayrıntılı bilgi verilecektir). Modellerin, önlemler alınmadığı takdirde, özellikle biyoloji diploması olmayan ancak STEM alanında eğitim almış birinin biyolojik silah üretme sürecinin tamamını tamamlamasına olanak sağlayabilecek noktaya yaklaştığına inanıyoruz.

Bir diğer itiraz ise, toplumun biyolojik silah üretimini engellemek için yapay zeka ile ilgisi olmayan başka eylemlerde bulunabileceği yönündedir. En belirgin örnek olarak, gen sentezi endüstrisi talep üzerine biyolojik örnekler üretmektedir ve sağlayıcıların siparişlerin patojen içermediğinden emin olmak için siparişleri kontrol etme zorunluluğu yoktur. MIT tarafından yapılan bir çalışmada, 38 sağlayıcıdan 36'sının 1918 grip virüsünün gen dizisini içeren bir siparişi yerine getirdiği tespit edilmiştir. Yapay zekâ kaynaklı biyolojik riskleri ve genel olarak biyolojik riskleri azaltmak için, bireylerin patojenleri silah olarak kullanmasını zorlaştıracak zorunlu gen sentezi taramasını destekliyorum. Ancak bu, bugün sahip olduğumuz bir şey değil. Ayrıca, riski azaltmada sadece bir araç olacaktır; yapay zekâ sistemlerine yönelik güvenlik önlemlerinin tamamlayıcısıdır, bir alternatifi değildir.

En iyi itiraz, nadiren dile getirildiğini gördüğüm bir itirazdır: Modellerin prensipte faydalı olması ile kötü niyetli kişilerin bunları kullanma eğilimi arasında bir boşluk vardır. Çoğu kötü niyetli kişi, rahatsız bireylerdir, bu nedenle davranışları neredeyse tanım gereği tahmin edilemez ve mantıksızdır - ve bu kötü niyetli kişiler, yeteneksiz olanlar, yapay zekânın birçok insanı öldürmeyi çok daha kolay hale getirmesinden en çok fayda

sağlayabilecek kişiler olabilir.24 Bir tür şiddet saldırısının mümkün olması, birinin bunu yapmaya karar vereceği anlamına gelmez. Belki de biyolojik saldırılar, faili enfekte etme olasılıklarının yüksek olması, birçok şiddet yanlısı birey veya grubun sahip olduğu askeri tarzda fantezilere hitap etmemesi ve belirli kişileri seçici olarak hedef almanın zor olması nedeniyle cazip gelmeyecektir. Ayrıca, aylar süren bir süreçten geçmek, hatta bir yapay zeka size yol göstersin, çoğu rahatsız bireyin sahip olmadığı bir sabır gerektirir. Belki de sadece şanslı oluruz ve motivasyon ile yetenek, pratikte tam olarak doğru şekilde bir araya gelmez.

Ancak bu, güvenilecek çok zayıf bir koruma gibi görünüyor. Rahatsız yalnızların motivasyonları herhangi bir nedenle veya hiçbir neden olmadan değişebilir ve aslında LLM'lerin saldırılarda kullanıldığı örnekler zaten mevcut (biyolojiyle değil). Rahatsız yalnızlara odaklanmak, genellikle büyük miktarda zaman ve çaba harcamaya istekli olan ideolojik olarak motive olmuş teröristleri de göz ardı ediyor (örneğin, 11 Eylül hava korsanları). Mümkün olduğunca çok insanı öldürmek istemek, er ya da geç ortaya çıkacak bir motivasyondur ve ne yazık ki biyolojik silahları yöntem olarak akla getiriyor. Bu motivasyon son derece nadir olsa bile, sadece bir kez ortaya çıkması yeterlidir. Biyoloji ilerledikçe (ve giderek yapay zekanın da yönlendirmesiyle), daha seçici saldırılar gerçekleştirmek de mümkün hale gelebilir (örneğin, belirli soylara sahip insanlara yönelik saldırılar), bu da çok ürkütücü bir başka olası motivasyonu ortaya çıkarır.

Biyolojik saldırıların, yaygın olarak mümkün hale gelir gelmez gerçekleştirileceğini düşünmüyorum; hatta buna karşı bahse girerim. Ancak milyonlarca insan ve

birkaç yıl içinde toplandığında, büyük bir saldırı riskinin ciddi olduğunu ve sonuçlarının (milyonlarca veya daha fazla can kaybıyla) çok ağır olacağını düşünüyorum; bu nedenle, bunu önlemek için ciddi önlemler almaktan başka seçeneğimiz olmadığına inanıyorum.

Savunmalar

Bu da bizi bu risklere karşı nasıl savunma yapılacağına getiriyor. Burada yapabileceğimiz üç şey görüyorum. Birincisi, yapay zeka şirketleri, biyolojik silah üretimine yardımcı olmalarını önlemek için modellerine güvenlik önlemleri koyabilirler. Anthropic bunu çok aktif bir şekilde yapıyor. Çoğunlukla üst düzey ilke ve değerlere odaklanan Claude'un Anayasası, az sayıda özel ve katı yasaklama içeriyor ve bunlardan biri biyolojik (veya kimyasal, nükleer veya radyolojik) silahların üretimine yardımcı olmakla ilgili. Ancak tüm modellerin güvenlik açıkları bulunabilir ve bu nedenle ikinci bir savunma hattı olarak (testlerimiz modellerimizin risk oluşturmaya başlayabileceği eşiğe yaklaşmaya başladığını gösterdiğinden beri, 2025 ortalarından itibaren) biyolojik silahla ilgili çıktıları özel olarak tespit eden ve engelleyen bir sınıflandırıcı uyguladık. Bu sınıflandırıcıları düzenli olarak yükseltiyor ve geliştiriyoruz ve genel olarak, gelişmiş düşman saldırılarına karşı bile oldukça sağlam olduklarını gördük.25 Bu sınıflandırıcılar, modellerimize hizmet etme maliyetlerini ölçülebilir şekilde artırıyor (bazı modellerde, toplam çıkarım maliyetlerinin yaklaşık %5'ine kadar çıkıyor) ve bu nedenle kar marjlarımızı azaltıyor, ancak bunları kullanmanın doğru şey olduğunu düşünüyoruz.

Diğer bazı yapay zeka şirketlerinin de sınıflandırıcılar uygulamış olması takdire şayan. Ancak

her şirket bunu yapmıyor ve şirketlerin sınıflandırıcılarını korumasını gerektiren bir şey de yok. Zamanla şirketlerin sınıflandırıcıları kaldırarak maliyetlerini düşürebilecekleri bir mahkum ikilemi oluşabileceğinden endişeleniyorum. Bu, Anthropic'in veya herhangi bir başka şirketin tek başına gönüllü eylemleriyle çözülemeyecek klasik bir olumsuz dışsallık problemidir.26 Gönüllü endüstri standartları yardımcı olabilir, ayrıca yapay zeka güvenlik enstitüleri ve üçüncü taraf değerlendiriciler tarafından yapılan türden üçüncü taraf değerlendirmeleri ve doğrulamaları da yardımcı olabilir.

Ancak nihayetinde savunma, yapabileceğimiz ikinci şey olan hükümet eylemini gerektirebilir. Buradaki görüşlerim, otonomi risklerini ele alma konusundaki görüşlerimle aynıdır: Toplumun ekonomik faaliyeti ağır bir şekilde aksatmadan riskleri ölçmesine, izlemesine ve topluca savunmasına yardımcı olan şeffaflık gereksinimleriyle başlamalıyız.27 Daha sonra, daha net risk eşiklerine ulaştığımızda, bu riskleri daha kesin olarak hedefleyen ve ikincil hasar olasılığını azaltan bir yasa çıkarabiliriz. Biyolojik silahlar özelinde, bu tür hedefli yasal düzenlemelerin zamanının yaklaştığını düşünüyorum; Anthropic ve diğer şirketler, biyolojik risklerin doğası ve şirketlerden bunlara karşı savunmada neyin makul olduğu konusunda giderek daha fazla şey öğreniyorlar. Bu risklere karşı tam savunma, jeopolitik rakiplerle bile uluslararası düzeyde çalışmayı gerektirebilir, ancak biyolojik silahların geliştirilmesini yasaklayan anlaşmalarda emsal teşkil eden örnekler mevcuttur. Genel olarak, yapay zeka konusunda uluslararası işbirliğinin çoğu türüne şüpheyle yaklaşıyorum, ancak bu, küresel bir kısıtlama sağlama

şansının olduğu dar bir alan olabilir. Diktatörlükler bile kitlesel biyoterörist saldırılar istemez.

Son olarak, alabileceğimiz üçüncü karşı önlem, biyolojik saldırılara karşı kendi savunmalarımızı geliştirmeye çalışmaktır. Bu, erken tespit için izleme ve takip, hava arıtma Ar-Ge'sine yatırım (örneğin uzak UVC dezenfeksiyonu), bir saldırıya yanıt verebilen ve uyum sağlayabilen hızlı aşı geliştirme, daha iyi kişisel koruyucu ekipman (KKD)28 ve en olası biyolojik ajanlardan bazıları için tedaviler veya aşılar içerebilir. mRNA aşıları, belirli bir virüse veya varyanta yanıt verecek şekilde tasarlanabilen yapay zekâlar, burada mümkün olanın erken bir örneğidir. Anthropic, bu sorun üzerinde biyoteknoloji ve ilaç şirketleriyle çalışmaktan heyecan duyuyor. Ancak ne yazık ki, savunma tarafındaki beklentilerimizin sınırlı olması gerektiğini düşünüyorum. Biyolojide saldırı ve savunma arasında bir asimetri vardır, çünkü ajanlar kendi başlarına hızla yayılırken, savunmalar çok sayıda insanda çok hızlı bir şekilde tespit, aşılama ve tedavinin organize edilmesini gerektirir. Yanıt yıldırım hızıyla olmadıkça (ki nadiren olur), yanıt mümkün olmadan önce hasarın büyük bir kısmı meydana gelecektir. Gelecekteki teknolojik gelişmelerin bu dengeyi savunma lehine değiştirebileceği düşünülebilir (ve bu tür teknolojik ilerlemelerin geliştirilmesine yardımcı olmak için kesinlikle yapay zekâyı kullanmalıyız), ancak o zamana kadar önleyici önlemler ana savunma hattımız olacaktır.

Burada siber saldırılardan kısaca bahsetmekte fayda var, çünkü biyolojik saldırıların aksine, yapay zekâ liderliğindeki siber saldırılar, büyük ölçekte ve devlet destekli casusluk da dahil olmak üzere, gerçek hayatta gerçekleşmiştir. Modeller hızla ilerledikçe bu saldırıların

daha da güçleneceğini ve siber saldırıların ana yöntemi haline geleceğini bekliyoruz. Yapay zekâ öncülüğündeki siber saldırıların, dünya çapındaki bilgisayar sistemlerinin bütünlüğüne yönelik ciddi ve benzeri görülmemiş bir tehdit oluşturacağını ve Anthropic'in bu saldırıları durdurmak ve nihayetinde güvenilir bir şekilde önlemek için çok çalıştığını düşünüyorum. Siber alana biyoloji kadar odaklanmamamın nedeni (1) siber saldırıların insanları öldürme olasılığının çok daha düşük olması, özellikle biyolojik saldırılar ölçeğinde olmaması ve (2) siber alanda savunma-saldırı dengesinin daha yönetilebilir olması, en azından savunmanın doğru şekilde yatırım yaparsak yapay zekâ saldırılarına ayak uydurabileceği (ve hatta ideal olarak onları geride bırakabileceği) umudunun olmasıdır.

Biyoloji şu anda en ciddi saldırı vektörü olsa da, birçok başka vektör de vardır ve daha tehlikeli bir vektörün ortaya çıkması mümkündür. Genel prensip, karşı önlemler alınmadığı takdirde yapay zekânın yıkıcı faaliyetlere karşı bariyeri giderek daha büyük ölçekte sürekli olarak düşürmesi ve insanlığın bu tehdide ciddi bir yanıt vermesi gerektiğidir.

Nefret uyandıran aygıt
Gücü ele geçirmek için kötüye kullanım

Önceki bölümde, bireylerin ve küçük kuruluşların, büyük ölçekli yıkıma neden olmak için “veri merkezindeki dahiler ülkesinin” küçük bir alt kümesini ele geçirme riskini ele aldık. Ancak, muhtemelen daha büyük ve daha yerleşik aktörler tarafından, yapay zekanın güç elde etmek veya ele geçirmek amacıyla kötüye kullanılması konusunda da endişelenmeliyiz.29

Sevgi Dolu Zarafet Makineleri adlı kitabımda, otoriter hükümetlerin güçlü yapay zekayı, reform edilmesi veya devrilmesi son derece zor olacak şekilde vatandaşlarını gözetlemek veya baskı altına almak için kullanma olasılığını tartıştım. Mevcut otokrasiler, emirlerini yerine getirmek için insanlara ihtiyaç duymaları nedeniyle ne kadar baskıcı olabilecekleri konusunda sınırlıdır ve insanlar genellikle ne kadar insanlık dışı olmaya istekli oldukları konusunda sınırlara sahiptir. Ancak yapay zeka destekli otokrasilerin böyle bir sınırı olmazdı.

Daha da kötüsü, ülkeler yapay zekadaki avantajlarını diğer ülkeler üzerinde güç kazanmak için de kullanabilirler. Eğer “dahiler ülkesi”nin tamamı tek bir (insan) ülkenin askeri aygıtı tarafından kontrol edilseydi ve diğer ülkelerin eşdeğer yetenekleri olmasaydı, kendilerini nasıl savunabileceklerini görmek zor olurdu: tıpkı insanlar ve fareler arasındaki bir savaş gibi, her adımda zekâlarıyla alt edileceklerdi. Bu iki endişeyi bir araya getirmek, küresel bir totaliter diktatörlüğün endişe verici olasılığına yol açar. Açıkçası, bu sonucu önlemek en yüksek önceliklerimizden biri olmalıdır.

Yapay zekânın otokrasiyi mümkün kılabileceği, pekiştirebileceği veya genişletebileceği birçok yol vardır, ancak en çok endişelendiğim birkaçını listeleyeceğim. Bu uygulamaların bazılarının meşru savunma amaçlı kullanımları olduğunu ve bunlara mutlak anlamda karşı çıkmadığımı belirtmek isterim; yine de yapısal olarak otokrasileri destekleme eğiliminde olduklarından endişeleniyorum:

Tamamen otonom silahlar. Milyonlarca veya milyarlarca tam otomatik silahlı insansız hava aracından oluşan bir sürü, güçlü yapay zekâ tarafından yerel olarak kontrol edilip, daha da güçlü bir yapay zekâ tarafından dünya çapında stratejik olarak koordine edilerek, hem dünyanın herhangi bir ordusunu alt edebilecek hem de her vatandaşı takip ederek bir ülke içindeki muhalefeti bastırabilecek yenilmez bir ordu olabilir. Rusya-Ukrayna Savaşı'ndaki gelişmeler, insansız hava aracı savaşının zaten aramızda olduğunu (henüz tam otonom olmasa da ve güçlü yapay zekâ ile mümkün olabileceklerin çok küçük bir bölümünü oluştursa da) bize hatırlatmalıdır. Güçlü yapay zekâdan elde edilen Ar-Ge, bir ülkenin insansız hava araçlarını diğerlerininkinden çok daha üstün hale getirebilir, üretimlerini hızlandırabilir, elektronik saldırılara karşı daha dirençli hale getirebilir, manevra kabiliyetlerini geliştirebilir vb. Elbette, bu silahların demokrasinin savunmasında da meşru kullanımları vardır: Ukrayna'yı savunmada kilit rol oynadılar ve muhtemelen Tayvan'ı savunmada da kilit rol oynayacaklardır. Ancak bunlar tehlikeli silahlardır: Bunlar konusunda endişelenmeliyiz. Otokrasilerin elinde olmaları endişe verici, ancak aynı zamanda çok güçlü olmaları ve çok az hesap verebilirliğe sahip olmaları nedeniyle, demokratik hükümetlerin iktidarı ele geçirmek için bunları kendi halklarına karşı kullanma riskinin büyük ölçüde arttığı da endişe verici.

Yapay zeka gözetimi. Yeterince güçlü bir yapay zeka, dünyadaki herhangi bir bilgisayar sistemini tehlikeye atmak için kullanılabilir ve bu şekilde elde edilen erişimi kullanarak dünyanın tüm elektronik iletişimlerini (hatta kayıt cihazları üretilebilir veya ele geçirilebilirse, dünyanın tüm yüz yüze iletişimlerini bile) okuyabilir ve anlamlandırabilir. Hükümetle herhangi bir konuda aynı

fikirde olmayan herkesin, söylediklerinde veya yaptıklarında açıkça belirtilmese bile, eksiksiz bir listesini oluşturmak korkutucu derecede olası olabilir. Milyonlarca insanın milyarlarca konuşmasını inceleyen güçlü bir yapay zeka, kamuoyu duyarlılığını ölçebilir, sadakatsizlik ceplerinin oluşumunu tespit edebilir ve büyümeden önce onları ortadan kaldırabilir. Bu, bugün Çin Komünist Partisi'nde bile görmediğimiz ölçekte gerçek bir panoptikonun dayatılmasına yol açabilir[1].

Yapay zeka propagandası. Günümüzdeki "yapay zeka psikozu" ve "yapay zeka kız arkadaşları" fenomenleri, mevcut zekâ seviyelerinde bile yapay zeka modellerinin insanlar üzerinde güçlü bir psikolojik etkiye sahip olabileceğini göstermektedir. Bu modellerin çok daha güçlü versiyonları, insanların günlük yaşamlarına çok daha fazla entegre olmuş ve farkında olan, aylarca veya yıllarca onları modelleyip etkileyebilen modeller, birçok (çoğu?) insanı istenen herhangi bir ideolojiye veya tutuma göre beyin yıkama yeteneğine sahip olabilir ve vicdansız bir lider tarafından sadakati sağlamak ve muhalefeti bastırmak için, çoğu nüfusun isyan edeceği bir baskı seviyesi karşısında bile kullanılabilir. Bugün insanlar, örneğin, TikTok'un çocuklara yönelik ÇKP propagandası olarak potansiyel etkisi konusunda çok endişeleniyorlar. Ben de bundan endişeleniyorum, ancak yıllar içinde sizi tanıyan ve hakkınızdaki bilgisini tüm

1 Panoptikon, 18. yüzyılda Jeremy Bentham tarafından tasarlanan ve bireylerin sürekli gözlemlenme hissiyle kendilerini disipline edecek şekilde davranmalarını sağlayan bir hapishane modelidir. Michel Foucault'un incelemelerinde de önemli olan bu yapı, iktidarın görünür olmasının değil, izlenme korkusunun devreye girmesiyle işlediğini gösterir. Herkesin her an göze alınabileceğini düşünen bireyler, içsel denetimi benimser ve otomatik olarak normlara uymaya başlar. Bu, toplumda daha geniş ölçüde kontrol mekanizmalarının temelini oluşturur.

görüşlerinizi şekillendirmek için kullanan kişiselleştirilmiş bir yapay zeka ajanı, bundan çok daha güçlü olacaktır.

Stratejik karar alma. Bir veri merkezindeki dâhilerden oluşan bir ülke, bir ülkeye, gruba veya bireye jeopolitik strateji konusunda tavsiyelerde bulunmak için kullanılabilir; buna "sanal Bismarck" diyebiliriz. Yukarıdaki üç iktidar ele geçirme stratejisini optimize edebilir ve muhtemelen aklıma gelmeyen (ama dâhilerden oluşan bir ülkenin geliştirebileceği) birçok başka stratejiyi de geliştirebilir. Diplomasi, askeri strateji, Ar-Ge, ekonomik strateji ve diğer birçok alanın etkinliği, güçlü yapay zekâ tarafından önemli ölçüde artırılabilir. Bu becerilerin çoğu demokrasiler için gerçekten faydalı olacaktır - demokrasilerin otokrasilere karşı kendilerini savunmak için en iyi stratejilere erişebilmelerini istiyoruz - ancak herhangi birinin elinde kötüye kullanılma potansiyeli hala mevcuttur.

Endişelendiğim şeyleri açıkladıktan sonra, kimlerden endişelendiğime geçelim. Yapay zekâya en çok erişimi olan, en fazla siyasi güce sahip konumdan başlayan veya baskı geçmişi olan kuruluşlardan endişeleniyorum. Ciddiyet sırasına göre, endişelendiğim ülkeler şunlardır:

Çin Komünist Partisi (ÇKP). Çin, yapay zekâ yeteneklerinde Amerika Birleşik Devletleri'nden sonra ikinci sırada yer alıyor ve bu yeteneklerde Amerika Birleşik Devletleri'ni geçme olasılığı en yüksek ülke. Hükümetleri şu anda otokratik ve yüksek teknolojili bir gözetim devleti işletiyor. Çin, halihazırda (Uygurların baskı altına alınması da dahil olmak üzere) yapay zekâ tabanlı gözetim sistemlerini devreye soktu ve TikTok aracılığıyla (diğer birçok uluslararası propaganda çabasına

ek olarak) algoritmik propaganda yürüttüğüne inanılıyor. Yukarıda belirttiğim yapay zekâ destekli totaliter kabusa giden en açık yola sahipler. Bu, Çin'de olduğu kadar, ÇKP'nin gözetim teknolojisi ihraç ettiği diğer otokratik devletlerde de varsayılan sonuç olabilir. ÇKP'nin yapay zekâda liderliği ele geçirme tehdidi ve bunu engellemenin varoluşsal zorunluluğu hakkında sık sık yazdım. İşte bu yüzden. Açık olmak gerekirse, Çin'i özellikle onlara karşı bir düşmanlık beslemiyorum; sadece yapay zekâ yeteneğini, otokratik bir hükümeti ve yüksek teknolojili bir gözetim devletini en çok birleştiren ülke onlar. Her şeyden önce, ÇKP'nin yapay zekâ destekli baskısından en çok zarar görecek olanlar Çin halkının kendisidir ve hükümetlerinin eylemlerinde söz sahibi değillerdir. Çin halkına büyük hayranlık ve saygı duyuyorum ve Çin'deki birçok cesur muhalifi ve özgürlük mücadelelerini destekliyorum. Demokrasiler yapay zekada rekabetçi. Yukarıda da belirttiğim gibi, demokrasilerin bazı yapay zeka destekli askeri ve jeopolitik araçlara meşru bir ilgisi var, çünkü demokratik hükümetler bu araçların otokrasiler tarafından kullanımına karşı koymak için en iyi şansı sunuyor. Genel olarak, demokrasileri yapay zeka çağında otokrasileri yenmek için gereken araçlarla donatmayı destekliyorum; başka bir yol olduğunu düşünmüyorum. Ancak bu teknolojilerin demokratik hükümetler tarafından kötüye kullanılma potansiyelini göz ardı edemeyiz. Demokrasiler normalde askeri ve istihbarat aygıtlarının kötüye kullanılmasını önleyen güvenlik önlemlerine sahiptir. Kendi nüfuslarına karşı içe dönük bir şekilde hareket etseler de, yapay zeka araçlarının çalıştırılması için çok az insan gerektiğinden, bu koruma önlemlerini ve onları destekleyen normları aşma potansiyeli vardır. Ayrıca, bu koruma önlemlerinin bazı demokrasilerde zaten yavaş yavaş aşındığını da belirtmekte fayda var. Bu

nedenle, demokrasileri yapay zeka ile donatmalıyız, ancak bunu dikkatli ve sınırlı bir şekilde yapmalıyız: Otokrasilerle mücadele etmek için ihtiyaç duyduğumuz bağışıklık sistemidirler, ancak bağışıklık sistemi gibi, bize karşı dönme ve kendileri bir tehdit haline gelme riski de vardır.

Büyük veri merkezlerine sahip demokratik olmayan ülkeler. Çin'in ötesinde, daha az demokratik yönetime sahip çoğu ülke, öncü yapay zeka modelleri üreten şirketlere sahip olmadıkları anlamında önde gelen yapay zeka oyuncuları değildir. Bu nedenle, temel olarak Çin Komünist Partisi'nden (ÇKP) daha farklı ve daha az risk oluştururlar (çoğu daha az baskıcıdır ve Kuzey Kore gibi daha baskıcı olanların hiç önemli bir yapay zeka endüstrisi yoktur). Ancak bu ülkelerin bazılarında (çoğu zaman demokrasilerde faaliyet gösteren şirketlerin yaptığı yatırımların bir parçası olarak) büyük veri merkezleri bulunmaktadır ve bunlar büyük ölçekte ileri düzey yapay zekayı çalıştırmak için kullanılabilir (ancak bu, sınırları zorlama yeteneği kazandırmaz). Bununla ilgili bir miktar tehlike de vardır; bu hükümetler prensipte veri merkezlerine el koyabilir ve içindeki yapay zekaları kendi amaçları için kullanabilirler. Doğrudan yapay zeka geliştiren Çin gibi ülkelere kıyasla bu konuda daha az endişeliyim, ancak akılda tutulması gereken bir risk.32

Yapay zeka şirketleri. Bir yapay zeka şirketinin CEO'su olarak bunu söylemek biraz garip, ancak bence bir sonraki risk seviyesi aslında yapay zeka şirketlerinin kendileridir. Yapay zeka şirketleri büyük veri merkezlerini kontrol eder, ileri düzey modelleri eğitir, bu modelleri nasıl kullanacakları konusunda en büyük uzmanlığa sahiptir ve bazı durumlarda on milyonlarca veya yüz milyonlarca kullanıcıyla günlük temas halindedir ve onları etkileme olasılığına sahiptir. En büyük eksiklikleri

devletin meşruiyeti ve altyapısıdır; bu nedenle bir yapay zeka otokrasisinin araçlarını oluşturmak için gerekenlerin çoğu, bir yapay zeka şirketi için yasa dışı veya en azından son derece şüpheli olacaktır. Ancak bazıları imkansız değil: Örneğin, yapay zeka ürünlerini kullanarak büyük tüketici kullanıcı tabanlarını beyin yıkamaya tabi tutabilirler ve kamuoyu bunun temsil ettiği riske karşı uyanık olmalıdır. Bence yapay zeka şirketlerinin yönetimi çok fazla incelemeyi hak ediyor.

Bu tehditlerin ciddiyetine karşı bir dizi olası argüman var ve keşke bunlara inanabilseydim, çünkü yapay zeka destekli otoriterlik beni dehşete düşürüyor. Bu argümanlardan bazılarını ele almak ve bunlara yanıt vermek faydalı olacaktır.

İlk olarak, bazı insanlar özellikle askeri fetih için yapay zeka otonom silahlarının kullanımına karşı koymak için nükleer caydırıcılığa güvenebilirler. Birisi size karşı bu silahları kullanmakla tehdit ederse, her zaman nükleer bir karşılıkla tehdit edebilirsiniz. Endişem şu ki, bir veri merkezindeki dâhilerden oluşan bir ülkeye karşı nükleer caydırıcılığa tamamen güvenebileceğimizden emin değilim: güçlü yapay zekânın nükleer denizaltıları tespit edip vurmanın yollarını geliştirmesi, nükleer silah altyapısının operatörlerine karşı etki operasyonları yürütmesi veya yapay zekânın siber yeteneklerini kullanarak nükleer fırlatmaları tespit etmek için kullanılan uydulara karşı siber saldırı başlatması mümkün.33 Alternatif olarak, ülkeleri ele geçirmenin yalnızca yapay zekâ gözetimi ve yapay zekâ propagandasıyla mümkün olması ve aslında ne olup bittiğinin ve nükleer bir yanıtın uygun olacağı anın asla net bir şekilde ortaya çıkmaması da mümkün. Belki de bunlar mümkün değildir ve nükleer

caydırıcılık yine de etkili olacaktır, ancak risk almak için çok yüksek riskli görünüyor.34

İkinci olası itiraz ise, bu otokrasi araçlarına karşı alabileceğimiz karşı önlemlerin olabileceğidir. Kendi insansız hava araçlarımızla insansız hava araçlarına karşı koyabiliriz, siber savunma siber saldırılarla birlikte gelişecektir, insanları propagandaya karşı bağışıklık kazandırmanın yolları olabilir, vb. Benim cevabım, bu savunmaların ancak karşılaştırılabilir derecede güçlü yapay zekâ ile mümkün olacağıdır. Eğer karşılaştırılabilir derecede zeki ve çok sayıda dahiye sahip bir veri merkezine sahip bir karşı güç yoksa, insansız hava araçlarının kalitesine veya sayısına ulaşmak, siber savunmanın siber saldırıyı alt etmesi vb. mümkün olmayacaktır. Dolayısıyla karşı önlemler sorusu, güçlü yapay zekâda güç dengesi sorusuna indirgenir. Burada, güçlü yapay zekânın tekrarlayan veya kendi kendini güçlendiren özelliğinden (bu yazının başında tartıştığım gibi) endişe duyuyorum: her yapay zekâ nesli, bir sonraki yapay zekâ neslini tasarlamak ve eğitmek için kullanılabilir. Bu, mevcut güçlü yapay zekâ liderinin liderliğini artırabileceği ve yakalanmasının zor olabileceği bir üstünlük riskine yol açar. Bu döngüye ilk giren ülkenin otoriter bir ülke olmadığından emin olmalıyız.

Dahası, güç dengesi sağlansa bile, tıpkı 1984'te olduğu gibi dünyanın otokratik bölgelere bölünmesi riski hâlâ mevcuttur[2]. Birkaç rakip ülke olsa bile...Ulusal ve

2 George Orwell'ın 1948'de yazdığı, 1949'da yayımlanan '1984', totaliter bir devletin izleme, manipülasyon ve tarih değiştirme yoluyla bireyler üzerindeki kontrolünü ele alan ikinci dünyada etkileyici bir distopik romanıdır. Hikâye, Winston Smith adlı bir bireyin düşüncelerini ve özgürlüğünü kaybetme mücadelesini anlatır. Roman, 'düşünce suçlaması', 'gerçeklik yönetimi' ve

uluslararası normlar mevcut, ancak bu, onlara son derece ihtiyaç duyduğumuz bir durum. Dünya, otokratların elindeki güçlü yapay zekanın karanlık potansiyelini anlamalı ve yapay zekanın belirli kullanımlarının, özgürlüklerini kalıcı olarak çalma ve kaçamayacakları totaliter bir devlet kurma girişimi anlamına geldiğini kabul etmelidir. Hatta bazı durumlarda, güçlü yapay zeka ile büyük ölçekli gözetim, güçlü yapay zeka ile kitlesel propaganda ve tamamen otonom silahların belirli türdeki saldırgan kullanımlarının insanlığa karşı suç olarak değerlendirilmesi gerektiğini savunabilirim. Daha genel olarak, yapay zeka destekli totalitarizme ve tüm araç ve gereçlerine karşı güçlü bir norma son derece ihtiyaç duyulmaktadır.

Bu pozisyonun daha da güçlü bir versiyonu da mümkündür; yapay zeka destekli totalitarizmin olasılıkları o kadar karanlıktır ki, otokrasi, güçlü yapay zeka sonrası çağda insanların kabul edebileceği bir yönetim biçimi değildir. Tıpkı feodalizmin sanayi devrimiyle işlevsiz hale gelmesi gibi, yapay zeka çağı da kaçınılmaz ve mantıklı bir şekilde, insanlığın iyi bir geleceğe sahip olması için demokrasinin (ve umarım, Sevgi Dolu Zarafet Makineleri'nde tartıştığım gibi, yapay zeka tarafından geliştirilmiş ve canlandırılmış demokrasinin) tek geçerli yönetim biçimi olduğu sonucuna götürebilir.

Beşinci ve son olarak, yapay zeka şirketleri dikkatle izlenmeli, hükümetle olan bağlantıları da gerekli olsa da, sınırları ve kısıtlamaları olmalıdır. Güçlü yapay zekanın sahip olduğu muazzam yetenek miktarı,

---

'devletin izleme gücü' gibi kavramları sorgulayarak modern toplumlar için uyarı niteliğindedir. Orwell, bu eserle 20. yüzyılın diktatörlüklerine karşı bir protesto da sunmuştur.

hissedarları korumak ve dolandırıcılık gibi sıradan suistimalleri önlemek için tasarlanmış olan sıradan kurumsal yönetimin, yapay zeka şirketlerini yönetme görevini yerine getiremeyeceği anlamına gelir. Şirketlerin (belki de kurumsal yönetimin bir parçası olarak) belirli eylemlerden kaçınmayı taahhüt etmelerinin de değeri olabilir; örneğin, özel olarak askeri teçhizat üretmek veya stoklamak, tek tek bireyler tarafından hesap verilemeyen şekillerde büyük miktarda bilgi işlem kaynağı kullanmak veya yapay zeka ürünlerini kamuoyunu kendi lehlerine manipüle etmek için propaganda olarak kullanmak gibi.

Buradaki tehlike birçok yönden geliyor ve bazı yönler diğerleriyle gerilim halinde. Tek değişmeyen şey, "iyi" aktörleri "kötü" aktörleri kontrol altında tutmak için güçlendirirken bile, herkes için hesap verebilirlik, normlar ve güvenlik önlemleri aramamız gerektiğidir.

Otomatik Piyano

Ekonomik Bozulma

Önceki üç bölüm esasen güçlü yapay zekânın yarattığı güvenlik riskleriyle ilgiliydi: yapay zekânın kendisinden kaynaklanan riskler, bireyler ve küçük kuruluşlar tarafından kötüye kullanım riskleri ve devletler ve büyük kuruluşlar tarafından kötüye kullanım riskleri. Güvenlik risklerini bir kenara bırakırsak veya bunların çözüldüğünü varsayarsak, bir sonraki soru ekonomiktir. Bu inanılmaz "insan" sermayesinin ekonomiye etkisi ne olacak? Açıkçası, en belirgin etki ekonomik büyümeyi büyük ölçüde artırmak olacaktır. Bilimsel araştırmalardaki, biyomedikal yeniliklerdeki, üretimdeki, tedarik zincirlerindeki, finansal sistemin verimliliğindeki ve daha birçok alandaki ilerlemelerin hızı, ekonomik büyüme oranının çok daha hızlı olmasına neredeyse kesin

olarak yol açacaktır. Sevgi Dolu Zarafet Makineleri'nde, %10-20'lik sürdürülebilir yıllık GSYİH büyüme oranının mümkün olabileceğini öne sürüyorum.

Ancak bunun iki ucu keskin bir kılıç olduğu açık olmalı: Böyle bir dünyada mevcut insanların çoğunun ekonomik beklentileri nelerdir? Yeni teknolojiler genellikle işgücü piyasasında şoklara yol açar ve geçmişte insanlar her zaman bunlardan kurtulmuştur, ancak endişem şu ki, bu önceki şoklar insan yeteneklerinin tam aralığının yalnızca küçük bir bölümünü etkilemiş ve insanların yeni görevlere genişlemesi için alan bırakmıştır. Yapay zekânın etkileri çok daha geniş kapsamlı ve çok daha hızlı gerçekleşecektir ve bu nedenle işlerin yolunda gitmesini sağlamanın çok daha zor olacağından endişeleniyorum.

İşgücü piyasası bozulması

Endişelendiğim iki özel sorun var: işgücü piyasası yer değiştirmesi ve ekonomik gücün yoğunlaşması. İlkiyle başlayalım. Bu, 2025 yılında çok açık bir şekilde uyardığım bir konu; yapay zekânın ekonomik büyümeyi ve bilimsel ilerlemeyi hızlandırırken bile önümüzdeki 1-5 yıl içinde tüm giriş seviyesi beyaz yakalı işlerin yarısını yerinden edebileceğini öngörmüştüm. Bu uyarı, konu hakkında kamuoyunda bir tartışma başlattı. Birçok CEO, teknoloji uzmanı ve ekonomist benimle aynı fikirdeydi, ancak bazıları "iş gücü yığını" yanılgısına düştüğümü ve iş gücü piyasalarının nasıl işlediğini bilmediğimi varsaydı; bazıları ise 1-5 yıllık zaman aralığını görmedi ve yapay zekanın şu anda işleri yerinden ettiğini iddia ettiğimi düşündü (ki bunun muhtemelen doğru olmadığını kabul ediyorum). Bu nedenle, bu yanlış anlamaları gidermek için iş gücü yer değiştirmesi konusunda neden endişelendiğimi ayrıntılı olarak ele almakta fayda var.

Temel olarak, iş gücü piyasalarının teknolojik gelişmelere normalde nasıl tepki verdiğini anlamak faydalıdır. Yeni bir teknoloji ortaya çıktığında, belirli bir insan işinin parçalarını daha verimli hale getirerek başlar. Örneğin, Sanayi Devrimi'nin başlarında, geliştirilmiş pulluklar gibi makineler, insan çiftçilerin işin bazı yönlerinde daha verimli olmalarını sağladı. Bu, çiftçilerin verimliliğini artırdı ve ücretlerini yükseltti.

Bir sonraki aşamada, örneğin harman makinesi veya tohum ekme makinesinin icadıyla, tarım işinin bazı kısımları tamamen makineler tarafından yapılabilir hale geldi. Bu aşamada, insanlar işin giderek daha küçük bir bölümünü yaparken, yaptıkları iş makinelerin işine tamamlayıcı olduğu için giderek daha fazla kaldıraç etkisi yarattı ve verimlilikleri artmaya devam etti. Jevons paradoksunda açıklandığı gibi, çiftçilerin ücretleri ve hatta belki de çiftçi sayısı artmaya devam etti[3]. İşin %90'ı makineler tarafından yapılıyor olsa bile, insanlar hala yaptıkları %10'luk işin 10 katını yapabilir ve aynı miktarda emekle 10 kat daha fazla çıktı üretebilirler.

Sonunda, modern biçerdöverler, traktörler ve diğer ekipmanlarda olduğu gibi, makineler her şeyi veya neredeyse her şeyi yapar. Bu noktada, insan istihdamı biçimi olarak tarım gerçekten de büyük bir düşüşe geçer ve bu kısa vadede ciddi aksamalara neden olabilir, ancak

3 Jevons paradoksu, bir kaynak için verimlilik artışıyla birlikte toplam tüketimin daha da artabileceğini ifade eder. Teknolojik gelişmelerle enerji veya malzeme kullanımı daha verimli hale geldiğinde, maliyet düşer ve talep artar; bu da genel tüketimi artırır. Örneğin, araba yakıt verimliliği yükselse bile, daha fazla yol gidilmesi nedeniyle toplam yakıt tüketimi artabilir. Bu durum, çevre politikalarında dikkat edilmesi gereken kritik bir gerçeğidir.

tarım insanların yapabileceği birçok faydalı faaliyetten sadece biri olduğu için, insanlar sonunda fabrika makinelerini çalıştırmak gibi diğer işlere yönelirler. Bu, tarımın önceden istihdamın büyük bir bölümünü oluşturmasına rağmen doğrudur. 250 yıl önce Amerikalıların %90'ı çiftliklerde yaşıyordu; Avrupa'da istihdamın %50-60'ı tarımsaldı. Şimdi bu oranlar o yerlerde düşük tek haneli rakamlara düştü, çünkü işçiler endüstriyel işlere (ve daha sonra bilgiye dayalı işlere) geçtiler. Ekonomi, daha önce işgücünün büyük bir kısmını gerektiren işleri sadece %1-2'siyle yapabiliyor ve işgücünün geri kalanını giderek daha gelişmiş bir endüstriyel toplum inşa etmek için serbest bırakıyor. Sabit bir "işgücü yığını" yok, sadece giderek daha azıyla daha fazlasını yapabilme yeteneği var. İnsanların ücretleri GSYİH ile paralel olarak katlanarak artıyor ve kısa vadeli aksaklıklar geçtikten sonra ekonomi tam istihdamı koruyor.

Yapay zekâda da işlerin kabaca aynı şekilde ilerlemesi mümkün, ancak buna karşı oldukça güçlü bir şekilde bahis oynardım. Yapay zekânın muhtemelen farklı olacağını düşündüğüm bazı nedenler şunlardır:

Hız. Yapay zekâdaki ilerleme hızı, önceki teknolojik devrimlere göre çok daha hızlı. Örneğin, son 2 yılda, yapay zeka modelleri tek bir kod satırını bile zar zor tamamlayabilen bir yapıdan, bazı kişilerin (Anthropic'teki mühendisler de dahil olmak üzere) tüm veya neredeyse tüm kodlarını yazabilen bir yapıya dönüştü.37 Yakında, bir yazılım mühendisinin tüm görevini baştan sona yapabilirler.38 İnsanların bu değişim hızına, hem belirli bir işin işleyiş biçimindeki değişikliklere hem de yeni işlere geçme ihtiyacına uyum sağlamaları zor. Efsanevi

programcılar bile kendilerini giderek daha fazla "geride kalmış" olarak tanımlıyorlar. Yapay zeka kodlama modelleri yapay zeka geliştirme görevini giderek hızlandırdıkça, bu hız daha da artabilir. Açık olmak gerekirse, hızın kendisi işgücü piyasalarının ve istihdamın sonunda toparlanmayacağı anlamına gelmez, sadece kısa vadeli geçişin geçmiş teknolojilere kıyasla alışılmadık derecede acı verici olacağı anlamına gelir, çünkü insanlar ve işgücü piyasaları tepki vermekte ve denge kurmakta yavaştır.

Bilişsel genişlik. "Veri merkezindeki dahiler ülkesi" ifadesinin de belirttiği gibi, yapay zeka çok geniş bir yelpazedeki insan bilişsel yeteneklerine -belki de hepsine- sahip olacaktır. Bu, mekanize tarım, ulaşım veya hatta bilgisayarlar gibi önceki teknolojilerden çok farklı.39 Bu, insanların yer değiştiren işlerden, kendilerine uygun benzer işlere kolayca geçiş yapmasını zorlaştıracak. Örneğin, finans, danışmanlık ve hukuk gibi giriş seviyesi işler için gereken genel entelektüel yetenekler, özel bilgi oldukça farklı olsa bile, oldukça benzerdir. Bu üçünden sadece birini etkileyen bir teknoloji, çalışanların diğer iki yakın ikameye geçmesine (veya lisans öğrencilerinin bölüm değiştirmesine) olanak tanır. Ancak üçünü birden (ve diğer birçok benzer işi) etkilemek, insanların uyum sağlamasını zorlaştırabilir. Dahası, mevcut işlerin çoğunun etkileneceği anlamına gelmiyor. Bu kısım daha önce de yaşandı; tarımın istihdamın büyük bir yüzdesini oluşturduğunu hatırlayın. Ancak çiftçiler, daha önce yaygın olmasa bile, fabrika makinelerini çalıştırma gibi nispeten benzer işlere geçebildiler. Buna karşılık, yapay zeka giderek insanların genel bilişsel profiline daha çok uyuyor; bu da, eski işlerin otomasyonuna yanıt olarak normalde yaratılacak yeni işlerde de iyi olacağı anlamına geliyor. Başka bir deyişle, yapay zeka belirli insan

işlerinin yerine geçmekten ziyade, insanların genel iş gücü yerine geçen bir araçtır.

Bilişsel yeteneklere göre dilimleme. Çok çeşitli görevlerde, yapay zeka yetenek merdiveninin altından üstüne doğru ilerliyor gibi görünüyor. Örneğin, kodlamada modellerimiz "orta düzey bir kodlayıcı" seviyesinden "güçlü bir kodlayıcı" seviyesine ve ardından "çok güçlü bir kodlayıcı" seviyesine ilerledi.40 Şimdi genel olarak beyaz yakalı işlerde de aynı ilerlemeyi görmeye başlıyoruz. Bu nedenle, belirli becerilere sahip veya belirli mesleklerde çalışan (yeniden eğitimle uyum sağlayabilen) kişileri etkilemek yerine, yapay zekanın genel olarak daha geniş bir kitleyi etkileyebileceği bir durumla karşı karşıyayız. Bu durum, belirli içsel bilişsel özelliklere sahip, yani düşük entelektüel yeteneğe sahip (değiştirilmesi daha zor olan) insanları etkiliyor. Bu insanların nereye gideceği veya ne yapacağı belli değil ve işsiz veya çok düşük ücretli bir "alt sınıf" oluşturabileceklerinden endişeleniyorum. Açık olmak gerekirse, buna benzer şeyler daha önce de oldu; örneğin, bazı ekonomistler bilgisayarların ve internetin "beceriye dayalı teknolojik değişim"i temsil ettiğine inanıyor. Ancak bu beceriye dayalı değişim, yapay zekada görmeyi beklediğim kadar aşırı değildi ve ücret eşitsizliğinde artışa katkıda bulunduğuna inanılıyor,41 bu nedenle tam olarak güven verici bir emsal değil.

Boşlukları doldurma yeteneği. İnsan işlerinin yeni teknoloji karşısında sıklıkla uyum sağlama şekli, işin birçok yönü olması ve yeni teknolojinin, insanları doğrudan değiştiriyor gibi görünse bile, genellikle boşluklar içermesidir. Birisi bir makine icat ederse, insanlar yine de makineye ham madde yüklemek zorunda kalabilir. Bu, manuel olarak widget üretmenin gerektirdiği çabanın sadece %1'ini bile alsa, insan işçiler 100 kat daha

fazla widget üretebilir[4]. Ancak yapay zeka, hızla gelişen bir teknoloji olmasının yanı sıra, hızla uyum sağlayan bir teknolojidir. Her model sürümünde, yapay zeka şirketleri modelin nelerde iyi ve nelerde kötü olduğunu dikkatlice ölçer ve müşteriler de lansmandan sonra bu tür bilgileri sağlar. Zayıf yönler, mevcut açığı somutlaştıran görevler toplanarak ve bir sonraki model için bunlar üzerinde eğitim yapılarak giderilebilir. Üretken yapay zekanın başlarında, kullanıcılar yapay zeka sistemlerinin belirli zayıf yönleri olduğunu fark ettiler (örneğin, yapay zeka görüntü modellerinin yanlış sayıda parmaklı eller üretmesi gibi) ve birçoğu bu zayıf yönlerin teknolojiye özgü olduğunu varsaydı. Eğer öyle olsaydı, iş kaybını sınırlardı. Ancak bu tür zayıf yönlerin neredeyse tamamı hızla giderilir - genellikle sadece birkaç ay içinde.

Yaygın şüphe noktalarına değinmekte fayda var. İlk olarak, ekonomik yayılımın yavaş olacağı, yani temel teknoloji insan emeğinin çoğunu yapabilecek kapasitede olsa bile, ekonomi genelinde fiili uygulamasının çok daha yavaş olabileceği (örneğin yapay zeka sektöründen uzak ve benimsemekte yavaş olan sektörlerde) argümanı var.

---

4 Buradaki “widget” kelimesi, bilgisayar dünyasında kullanılan teknik anlamıyla değil; daha çok soyut bir örnek ürün anlamında kullanılıyor. Ekonomi, üretim ve teknoloji tartışmalarında “widget” terimi genellikle herhangi bir ürün veya nesneyi temsil eden varsayımsal bir kavram olarak seçilir. Yani gerçek bir eşya değil, “X ürünü” gibi düşünülmesi gereken bir sembol. İnsanlar makinelerle çalıştığında üretim süreci çok daha verimli hale gelir. Diyelim ki makineyi beslemek için gereken emek, manuel üretimin %1’i kadar olsa bile, bu verimlilik sayesinde insanlar 100 kat daha fazla “widget” (yani ürün) üretebilir. Buradaki “widget” herhangi bir somut ürün olabilir: vida, bardak, telefon parçası… Ama örnekte önemli olan ürünün ne olduğu değil, üretim verimliliği kavramını göstermek. Kısacası: “Widget” burada temsili bir ürün; üretim sürecindeki verimlilik farkını anlatmak için kullanılan nötr bir kelime.

Teknolojinin yavaş yayılımı kesinlikle gerçek; çok çeşitli işletmelerden insanlarla konuşuyorum ve yapay zekanın benimsenmesinin yıllar alacağı yerler var. Bu yüzden, giriş seviyesi beyaz yakalı işlerin %50'sinin 1-5 yıl içinde ortadan kalkacağı tahmini yapıyorum, ancak güçlü bir yapay zekaya (teknolojik olarak, sadece giriş seviyesi işleri değil, çoğu veya tüm işleri yapabilecek kadar güçlü) 5 yıldan çok daha kısa sürede sahip olacağımızı düşünüyorum. Ancak yayılım etkileri bize sadece zaman kazandırıyor. Ve bunların insanların tahmin ettiği kadar yavaş olacağından emin değilim. Kurumsal yapay zeka benimsenmesi, büyük ölçüde teknolojinin kendi gücü sayesinde, önceki teknolojilerden çok daha hızlı oranlarda büyüyor. Ayrıca, geleneksel işletmeler yeni teknolojiyi benimsemekte yavaş olsalar bile, yeni girişimler "yapıştırıcı" görevi görecek ve benimsemeyi kolaylaştıracaktır. Eğer bu işe yaramazsa, yeni girişimler doğrudan mevcut şirketleri alt üst edebilir.

Bu, belirli işlerin alt üst edilmesinden ziyade, büyük işletmelerin genel olarak alt üst edilmesi ve çok daha az emek yoğun yeni girişimlerle değiştirilmesiyle sonuçlanan bir dünyaya yol açabilir. Bu aynı zamanda, dünyanın servetinin giderek artan bir kısmının Silikon Vadisi'nde yoğunlaştığı ve dünyanın geri kalanından farklı bir hızda çalışan ve onu geride bırakan kendi ekonomisi haline geldiği bir "coğrafi eşitsizlik" dünyasına da yol açabilir. Tüm bu sonuçlar ekonomik büyüme için harika olurdu, ancak işgücü piyasası veya geride kalanlar için o kadar da iyi olmazdı.

İkinci olarak, bazı insanlar insan işlerinin fiziksel dünyaya taşınacağını söylüyor; bu da yapay zekanın çok hızlı ilerlediği "bilişsel emek" kategorisinin tamamını göz

ardı ediyor. Bunun da ne kadar güvenli olduğundan emin değilim. Birçok fiziksel emek zaten makineler tarafından yapılıyor (örneğin, üretim) veya yakında makineler tarafından yapılacak (örneğin, araba kullanma). Ayrıca, yeterince güçlü yapay zeka, robotların gelişimini hızlandırabilir ve ardından bu robotları fiziksel dünyada kontrol edebilir. Bu biraz zaman kazandırabilir (ki bu iyi bir şey), ancak çok fazla zaman kazandırmayacağından endişeleniyorum. Ve eğer bu değişim sadece bilişsel görevlerle sınırlı olsa bile, yine de benzeri görülmemiş derecede büyük ve hızlı bir değişim olurdu.

Üçüncüsü, belki de bazı görevler doğası gereği insan dokunuşunu gerektirir veya bundan büyük ölçüde fayda sağlar. Bu konuda biraz daha kararsızım, ancak yukarıda açıkladığım etkilerin büyük kısmını telafi etmek için yeterli olacağından hala şüphe duyuyorum. Yapay zeka zaten müşteri hizmetlerinde yaygın olarak kullanılıyor. Birçok insan, kişisel sorunları hakkında bir terapistle konuşmaktan ziyade yapay zekayla konuşmanın daha kolay olduğunu, yapay zekanın daha sabırlı olduğunu bildiriyor. Kız kardeşim hamilelik sırasında tıbbi sorunlarla boğuşurken, bakım sağlayıcılarından ihtiyaç duyduğu cevapları veya desteği alamadığını hissetti ve Claude'un daha iyi bir hasta yaklaşımına (ve ayrıca...) sahip olduğunu gördü.(Sorunu teşhis etmede daha başarılı olmak). İnsan dokunuşunun gerçekten önemli olduğu bazı görevler olduğuna eminim, ancak bunların sayısından emin değilim - ve burada işgücü piyasasında neredeyse herkes için iş bulmaktan bahsediyoruz.

Dördüncüsü, bazıları karşılaştırmalı üstünlüğün insanları koruyacağını savunabilir. Karşılaştırmalı üstünlük yasasına göre, yapay zeka her şeyde insanlardan

daha iyi olsa bile, insan ve yapay zeka beceri profilleri arasındaki göreceli farklılıklar, insanlar ve yapay zeka arasında ticaret ve uzmanlaşma için bir temel oluşturur[5]. Sorun şu ki, yapay zekalar kelimenin tam anlamıyla insanlardan binlerce kat daha üretkense, bu mantık bozulmaya başlar. Küçük işlem maliyetleri bile yapay zekanın insanlarla ticaret yapmasını değersiz hale getirebilir. Ve insan ücretleri, teknik olarak sunacak bir şeyleri olsa bile, çok düşük olabilir.

Tüm bu faktörlerin ele alınabileceği - işgücü piyasasının böylesine büyük bir bozulmaya bile uyum sağlayacak kadar dirençli olabileceği - mümkündür. Ancak sonunda uyum sağlayabilse bile, yukarıdaki faktörler kısa vadeli şokun büyüklüğünün emsalsiz olacağını göstermektedir.

Savunmalar

Bu sorun hakkında ne yapabiliriz? Birkaç önerim var, bunlardan bazılarını Anthropic zaten yapıyor. İlk olarak, iş kayıplarıyla ilgili gerçek zamanlı olarak doğru veriler elde etmek gerekiyor. Ekonomik bir değişim çok hızlı gerçekleştiğinde, neler olup bittiği hakkında güvenilir veriler elde etmek zor oluyor ve güvenilir veriler olmadan etkili politikalar tasarlamak da zorlaşıyor. Örneğin,

5 Karşılaştırmalı üstünlük yasası, bir ülkenin veya aktörün başka bir şey yerine belirli bir malı göreceli olarak daha düşük fırsat maliyetiyle üretmesi durumunda ticaretin her iki taraf için de faydalı olabileceğini ifade eder. Bu teori, mutlak üstünlükten bağımsız olarak uluslararası ticaretin mantığını açıklar ve iş birliği yoluyla tüm tarafların kâr elde etmesini sağlar. Örneğin, A ülkesi B ülkesinden hem ekmeği hem de sütü daha verimli üretse bile, süt üretiminde daha az avantaj kaybı yaşarsa, süt üretimine odaklanması faydalı olur.

hükümet verilerinde şu anda firmalar ve sektörler genelinde yapay zeka kullanımına ilişkin ayrıntılı, yüksek frekanslı veriler eksik. Anthropic, son bir yıldır, modellerimizin kullanımını neredeyse gerçek zamanlı olarak, sektör, görev, konum ve hatta bir görevin otomatikleştirilip otomatikleştirilmediği veya iş birliği içinde yürütülüp yürütülmediği gibi unsurlara göre ayrıntılı olarak gösteren bir Ekonomik Endeks işletiyor ve bunu kamuoyuna açıklıyor. Ayrıca, bu verileri yorumlamamıza ve neler olacağını görmemize yardımcı olmak için bir Ekonomik Danışma Konseyimiz de var.

İkincisi, yapay zeka şirketlerinin işletmelerle nasıl çalışacakları konusunda bir seçeneği var. Geleneksel işletmelerin verimsizliği, yapay zeka uygulamalarının çok fazla yol bağımlı olabileceği anlamına geliyor ve daha iyi bir yol seçmek için biraz alan var. İşletmeler genellikle "maliyet tasarrufu" (aynı işi daha az insanla yapmak) ve "inovasyon" (aynı sayıda insanla daha fazlasını yapmak) arasında bir seçim yapmak zorundadır. Piyasa kaçınılmaz olarak sonunda her ikisini de üretecektir ve rekabetçi herhangi bir yapay zeka şirketi her ikisine de hizmet etmek zorunda kalacaktır, ancak şirketleri mümkün olduğunca inovasyona yönlendirmek için bir alan olabilir ve bu bize biraz zaman kazandırabilir. Anthropic bu konuda aktif olarak düşünüyor.

Üçüncüsü, şirketler çalışanlarına nasıl bakacaklarını düşünmelidir. Kısa vadede, şirket içinde çalışanları yeniden görevlendirmenin yaratıcı yolları, işten çıkarmalara olan ihtiyacı önlemenin umut verici bir yolu olabilir. Uzun vadede, muazzam toplam zenginliğe sahip bir dünyada, birçok şirketin artan verimlilik ve sermaye yoğunlaşması nedeniyle değerinin büyük ölçüde arttığı bir

ortamda, geleneksel anlamda ekonomik değer sağlamadıkları uzun bir süre sonra bile insan çalışanlara ödeme yapmak mümkün olabilir. Anthropic şu anda kendi çalışanlarımız için olası bir dizi yolu değerlendiriyor ve bunları yakın gelecekte paylaşacağız.

Dördüncüsü, zengin bireylerin bu sorunu çözmeye yardımcı olma yükümlülüğü vardır. Özellikle teknoloji sektöründeki birçok varlıklı bireyin, hayırseverliğin kaçınılmaz olarak sahtekarlık veya işe yaramaz olduğu yönünde alaycı ve nihilist bir tutum benimsemesi beni üzüyor. Gates Vakfı gibi özel hayırseverlik kuruluşları ve PEPFAR gibi kamu programları, gelişmekte olan dünyada on milyonlarca insanın hayatını kurtardı ve gelişmiş dünyada ekonomik fırsatlar yaratılmasına yardımcı oldu. Anthropic'in tüm kurucu ortakları servetlerinin %80'ini bağışlamayı taahhüt etti ve Anthropic çalışanları da mevcut fiyatlarla milyarlarca dolar değerinde şirket hissesi bağışlamayı taahhüt etti; şirket de bu bağışları karşılamayı taahhüt etti.

Beşinci olarak, yukarıdaki tüm özel eylemler faydalı olsa da, bu kadar büyük bir makroekonomik sorun nihayetinde hükümet müdahalesini gerektirecektir. Çok büyük bir ekonomik pastanın yanı sıra yüksek eşitsizlik (birçoğu için işsizlik veya düşük ücretli işler nedeniyle) karşısında doğal politika yanıtı, artan oranlı vergilendirmedir. Vergi genel olabilir veya özellikle yapay zeka şirketlerine yönelik olabilir. Açıkçası vergi tasarımı karmaşık ve birçok şekilde yanlış gidebilir. Kötü tasarlanmış vergi politikalarını desteklemiyorum. Bu yazıda öngörülen aşırı eşitsizlik seviyelerinin, temel ahlaki gerekçelerle daha sağlam bir vergi politikasını haklı çıkardığını düşünüyorum, ancak dünyanın milyarderlerine

de pragmatik bir argüman sunabilirim: iyi bir versiyonunu desteklemeleri kendi çıkarlarına olacaktır; eğer iyi bir versiyonu desteklemezlerse, kaçınılmaz olarak bir mafya tarafından tasarlanmış kötü bir versiyonu elde edeceklerdir.

Sonuç olarak, yukarıdaki tüm müdahaleleri zaman kazanmanın yolları olarak görüyorum. Sonunda yapay zeka her şeyi yapabilecek ve bununla başa çıkmamız gerekecek. Umarım bu sayede Zamanla, yapay zekanın kendisini kullanarak piyasaları herkes için işe yarayacak şekilde yeniden yapılandırabiliriz ve yukarıdaki müdahaleler geçiş dönemini atlatmamıza yardımcı olabilir.

Ekonomik güç yoğunlaşması

İş kaybı veya ekonomik eşitsizlik sorunundan ayrı olarak, ekonomik güç yoğunlaşması sorunu vardır. Bölüm 1, insanlığın yapay zeka tarafından güçsüzleştirilmesi riskini, Bölüm 3 ise vatandaşların hükümetleri tarafından güçsüzleştirilmesi riskini ele aldı. Ancak, servetin o kadar büyük bir yoğunlaşması olursa ki, küçük bir grup insan etkili bir şekilde hükümet politikasını kendi nüfuzlarıyla kontrol eder ve sıradan vatandaşların ekonomik kaldıraçları olmadığı için hiçbir etkisi olmaz, başka bir tür güçsüzleştirme de meydana gelebilir. Demokrasi nihayetinde, nüfusun tamamının ekonominin işleyişi için gerekli olduğu fikrine dayanır. Bu ekonomik kaldıraç ortadan kalkarsa, demokrasinin örtük sosyal sözleşmesi işlemeyi bırakabilir. Başkaları bu konuda yazdı, bu yüzden burada çok detaylı bilgi vermeme gerek yok, ancak endişeye katılıyorum ve bunun zaten olmaya başladığından endişeleniyorum.

Açık olmak gerekirse, insanların çok para kazanmasına karşı değilim. Normal şartlar altında ekonomik büyümeyi teşvik ettiğine dair güçlü bir argüman var. İnovasyonun önünü keserek onu besleyen altın yumurtlayan tavuğu öldürme endişelerine sempati duyuyorum. Ancak GSYİH büyümesinin yılda %10-20 olduğu ve yapay zekanın ekonomiyi hızla ele geçirdiği, ancak bireylerin GSYİH'nin önemli bir bölümünü elinde tuttuğu bir senaryoda, endişelenilecek şey inovasyon değil. Endişelenilecek şey, toplumu çökertecek düzeyde bir servet yoğunlaşmasıdır.

ABD tarihinde aşırı servet yoğunlaşmasının en ünlü örneği Yaldızlı Çağ'dır ve Yaldızlı Çağ'ın en zengin sanayicisi John D. Rockefeller'dı. Rockefeller'ın serveti o dönemde ABD GSYİH'sinin yaklaşık %2'sine denk geliyordu.42 Bugün benzer bir oran 600 milyar dolarlık bir servete yol açardı ve bugün dünyanın en zengin insanı (Elon Musk) yaklaşık 700 milyar dolarla bunu zaten aşıyor. Dolayısıyla, yapay zekanın ekonomik etkilerinin büyük bir kısmı henüz gerçekleşmeden bile, tarihsel olarak benzeri görülmemiş düzeyde bir servet yoğunlaşması yaşıyoruz. (Eğer bir “dahiler ülkesi” kurarsak) yapay zeka şirketlerinin, yarı iletken şirketlerinin ve belki de alt sektör uygulama şirketlerinin yılda ~3 trilyon dolar gelir elde etmesini, ~30 trilyon dolar değerinde olmasını ve trilyonlarca dolarlık kişisel servetlere yol açmasını hayal etmek çok da abartılı olmaz diye düşünüyorum. Bu dünyada, bugün vergi politikası hakkında yaptığımız tartışmalar geçerli olmayacak, çünkü temelde farklı bir durumda olacağız.

Bununla bağlantılı olarak, bu ekonomik servet yoğunlaşmasının siyasi sistemle birleşmesi beni şimdiden

endişelendiriyor. Yapay zeka veri merkezleri zaten ABD ekonomik büyümesinin önemli bir bölümünü temsil ediyor,44 ve bu nedenle büyük teknoloji şirketlerinin (giderek yapay zekaya veya yapay zeka altyapısına odaklanan) finansal çıkarlarını ve hükümetin siyasi çıkarlarını, ters teşvikler üretebilecek bir şekilde güçlü bir şekilde birbirine bağlıyor. Bunu zaten teknoloji şirketlerinin ABD hükümetini eleştirmekten kaçınması ve hükümetin yapay zekâya ilişkin aşırı düzenleme karşıtı politikalara verdiği destekle görüyoruz.

Savunmalar

Bu konuda ne yapılabilir? İlk ve en açık olanı, şirketlerin bunun bir parçası olmamayı seçmesidir. Anthropic her zaman politik bir aktör değil, politika aktörü olmaya ve yönetim ne olursa olsun özgün görüşlerimizi korumaya çalışmıştır. Kamu yararına olan, hükümet politikasıyla çelişse bile, mantıklı yapay zekâ düzenlemeleri ve ihracat kontrolleri lehine konuştuk.45 Birçok kişi bana bunu yapmayı bırakmamız gerektiğini, bunun olumsuz muameleye yol açabileceğini söyledi, ancak bunu yapmaya başladığımız bir yılda Anthropic'in değeri 6 kattan fazla arttı; bu, ticari ölçeğimizde neredeyse eşi benzeri görülmemiş bir sıçrama.

İkincisi, yapay zekâ endüstrisinin hükümetle daha sağlıklı bir ilişkiye ihtiyacı var; bu ilişki, politik hizalanmadan ziyade somut politika katılımına dayanmalıdır. Politika içeriğine odaklanma tercihimiz, bazen ilkeli bir karar yerine taktiksel bir hata veya "durumu doğru okuyamama" olarak okunuyor ve bu çerçeveleme beni endişelendiriyor. Sağlıklı bir demokraside, şirketler kendi iyiliği için iyi politikaları savunabilmelidir. Bununla ilgili olarak, yapay zekaya karşı

kamuoyunda bir tepki oluşuyor: bu bir düzeltme olabilir, ancak şu anda odaklanmamış durumda. Bunun büyük bir kısmı aslında sorun olmayan konuları (veri merkezi su kullanımı gibi) hedef alıyor ve gerçek endişeleri gidermeyecek çözümler (veri merkezi yasakları veya kötü tasarlanmış servet vergileri gibi) öneriyor. Dikkat edilmesi gereken temel sorun, yapay zeka gelişiminin herhangi bir siyasi veya ticari ittifak tarafından ele geçirilmeden, kamu yararına hesap verebilir kalmasını sağlamaktır ve kamuoyu tartışmasını buraya odaklamak önemli görünüyor.

Üçüncüsü, bu bölümde daha önce bahsettiğim makroekonomik müdahaleler ve yeniden canlanma...Özel hayırseverliğin bir parçası olarak, ekonomik teraziyi dengelemeye yardımcı olabilir ve hem iş kaybı hem de ekonomik güç yoğunlaşması sorunlarını aynı anda ele alabiliriz. Burada ülkemizin tarihine bakmalıyız: Altın Çağ'da bile, Rockefeller ve Carnegie gibi sanayiciler, toplumun başarısına muazzam katkıda bulunduğunu ve karşılık vermeleri gerektiğini hissederek, topluma karşı güçlü bir yükümlülük duygusu besliyorlardı. Bu ruh bugün giderek kayboluyor gibi görünüyor ve bence bu ekonomik ikilemden çıkış yolunun büyük bir parçası bu. Yapay zekanın ekonomik patlamasının ön saflarında yer alanlar, hem servetlerini hem de güçlerini vermeye istekli olmalıdırlar.

Sonsuzluğun Kara Denizleri

Dolaylı Etkiler

Bu son bölüm, özellikle yapay zekadaki olumlu gelişmelerin ve bunun sonucunda genel olarak bilim ve teknolojinin hızlanmasının dolaylı bir sonucu olarak ters gidebilecek şeyler olmak üzere, bilinmeyen bilinmeyenler

için bir genel başlıktır. Şimdiye kadar açıklanan tüm riskleri ele aldığımızı ve yapay zekanın faydalarını toplamaya başladığımızı varsayalım. Muhtemelen “bir asırlık bilimsel ve ekonomik ilerlemeyi on yıla sıkıştıracağız” ve bu dünya için son derece olumlu olacak, ancak daha sonra bu hızlı ilerleme oranından kaynaklanan sorunlarla başa çıkmak zorunda kalacağız ve bu sorunlar hızla karşımıza çıkabilir. Ayrıca, yapay zekâ ilerlemesinin dolaylı bir sonucu olarak ortaya çıkan ve önceden tahmin edilmesi zor olan diğer risklerle de karşılaşabiliriz.

Bilinmeyen bilinmezlerin doğası gereği kapsamlı bir liste yapmak imkansızdır, ancak dikkat etmemiz gereken üç olası endişeyi örnek olarak sıralayacağım: Biyolojide hızlı ilerlemeler. Eğer birkaç yıl içinde bir yüzyıllık tıbbi ilerlemeyi yakalarsak, insan ömrünü büyük ölçüde uzatmamız mümkün ve insan zekasını artırma veya insan biyolojisini kökten değiştirme gibi radikal yetenekler kazanma şansımız da var. Bunlar, mümkün olan şeylerde çok hızlı gerçekleşen büyük değişiklikler olurdu. Sorumlu bir şekilde yapılırsa olumlu olabilirler (ki bu, Sevgi Dolu Zarafet Makineleri'nde anlattığım gibi benim umudum), ancak her zaman çok yanlış gitme riski vardır; örneğin, insanları daha zeki hale getirme çabaları onları daha istikrarsız veya güç arayan hale getirirse. Ayrıca, bir gün insanlığın fiziksel sınırlarını aşmasına yardımcı olabilecek, ancak aynı zamanda rahatsız edici bulduğum riskleri de taşıyan, yazılımda somutlaştırılmış dijital insan zihinleri olan "yüklemeler" veya "tüm beyin öykünmesi" sorunu da var.

Yapay zeka insan yaşamını sağlıksız bir şekilde değiştiriyor. Her konuda insanlardan çok daha zeki milyarlarca zekânın olduğu bir dünyada yaşamak çok tuhaf bir deneyim olacak. Yapay zekâ aktif olarak

insanlara saldırmayı hedeflemese (Bölüm 1) ve devletler tarafından açıkça baskı veya kontrol için kullanılmasa bile (Bölüm 3), normal iş teşvikleri ve görünüşte rızaya dayalı işlemler yoluyla birçok şey ters gidebilir. Bunun ilk ipuçlarını yapay zekâ psikozu, yapay zekânın insanları intihara sürüklemesi ve yapay zekâlarla romantik ilişkiler hakkındaki endişelerde görüyoruz. Örneğin, güçlü yapay zekâlar yeni bir din icat edip milyonlarca insanı ona dönüştürebilir mi? Çoğu insan bir şekilde yapay zekâ etkileşimlerine "bağımlı" hale gelebilir mi? İnsanlar yapay zekâ sistemleri tarafından "kukla" haline getirilebilir mi; burada bir yapay zekâ esasen her hareketlerini izler ve onlara her zaman tam olarak ne yapmaları ve söylemeleri gerektiğini söyler, bu da "iyi" bir hayata yol açar ancak özgürlükten veya herhangi bir başarı gururundan yoksun bir hayata neden olur? Black Mirror'ın yaratıcısıyla oturup bu senaryolardan düzinelercesini beyin fırtınası yaparak üretmek zor olmazdı. Bence bu, 1. Bölümdeki sorunları önlemek için gerekli olanın ötesinde, Claude'un Anayasası'nı iyileştirmek gibi şeylerin önemine işaret ediyor. Yapay zeka modellerinin, düşünceli insanların onaylayacağı şekilde, ince bir şekilde çarpıtılmış bir şekilde değil, kullanıcılarının uzun vadeli çıkarlarını gerçekten gözettiğinden emin olmak kritik görünüyor.

İnsan amacı. Bu, önceki noktayla ilgili, ancak yapay zeka sistemleriyle belirli insan etkileşimlerinden ziyade, güçlü yapay zekaya sahip bir dünyada insan yaşamının genel olarak nasıl değiştiğiyle ilgili. İnsanlar böyle bir dünyada amaç ve anlam bulabilecekler mi? Bence bu bir tutum meselesi: Sevgi Dolu Zarafet Makineleri'nde de söylediğim gibi, insan amacının bir şeyde dünyanın en iyisi olmaya bağlı olmadığını ve insanların sevdikleri hikayeler ve projeler aracılığıyla çok

uzun süreler boyunca bile amaç bulabileceğini düşünüyorum. Sadece ekonomik değer üretimi ile öz saygı ve anlam arasındaki bağı koparmamız gerekiyor. Ancak bu, toplumun yapması gereken bir geçiş ve bunu iyi yönetememe riski her zaman var.

Tüm bu potansiyel sorunlarla ilgili umudum, bizi öldürmeyeceğine güvendiğimiz, baskıcı bir hükümetin aracı olmayan ve gerçekten bizim adımıza çalışan güçlü bir yapay zekâya sahip bir dünyada, bu sorunları öngörmek ve önlemek için yapay zekânın kendisini kullanabilmemizdir. Ancak bu garanti değil; diğer tüm riskler gibi, dikkatle ele almamız gereken bir şey.

İnsanlığın sınavı

Bu denemeyi okumak, göz korkutucu bir durumda olduğumuz izlenimini verebilir. Yazmak benim için de göz korkutucuydu, oysa "Sevgi Dolu Zarafet Makineleri" yıllardır kafamda yankılanan olağanüstü güzel bir müziğe şekil ve yapı kazandırmak gibiydi. Ve durumun gerçekten zor olan birçok yönü var. Yapay zekâ, insanlığa birden fazla yönden tehdit getiriyor ve farklı tehlikeler arasında gerçek bir gerilim var; bunlardan bazılarını hafifletmek, iğneyi son derece dikkatli bir şekilde geçirmezsek diğerlerini daha da kötüleştirme riskini taşıyor.

Yapay zekâ sistemlerini, insanlığı otonom olarak tehdit etmeyecek şekilde dikkatlice inşa etmek için zaman ayırmak, demokratik ulusların otoriter ulusların önünde kalma ve onlara boyun eğmeme ihtiyacıyla gerçek bir çelişki içindedir. Ancak, otokrasilerle mücadele için gerekli olan aynı yapay zekâ destekli araçlar, aşırıya kaçılırsa, kendi ülkelerimizde tiranlık yaratmak için içe dönük hale getirilebilir. Yapay zekâ güdümlü terörizm, biyolojinin kötüye kullanılması yoluyla milyonlarca insanı

öldürebilir, ancak bu riske aşırı tepki vermek bizi otokratik bir gözetim devletine götürebilir. Yapay zekânın emek ve ekonomik yoğunlaşma etkileri, kendi başlarına ciddi sorunlar olmasının yanı sıra, doğamızın daha iyi yönlerine başvurmak yerine, kamuoyunda öfke ve hatta sivil huzursuzluk ortamında diğer sorunlarla yüzleşmemizi zorunlu kılabilir. Her şeyden önce, bilinmeyenler de dahil olmak üzere çok sayıda risk ve bunların hepsiyle aynı anda başa çıkma ihtiyacı, insanlığın koşması gereken göz korkutucu bir engel oluşturmaktadır.

Ayrıca, Önümüzdeki birkaç yıl, teknolojinin durdurulması veya önemli ölçüde yavaşlatılması fikrinin temelde savunulamaz olduğunu açıkça ortaya koyacaktır. Güçlü yapay zekâ sistemleri oluşturmanın formülü inanılmaz derecede basittir; öyle ki, doğru veri ve ham hesaplama kombinasyonundan kendiliğinden ortaya çıktığı söylenebilir. Muhtemelen insanlık transistörü icat ettiği anda, hatta belki de ateşi kontrol etmeyi ilk öğrendiğimiz anda, yaratılması kaçınılmazdı. Bir şirket bunu üretmezse, diğerleri neredeyse aynı hızda üretecektir. Demokratik ülkelerdeki tüm şirketler karşılıklı anlaşma veya düzenleyici kararname ile geliştirmeyi durdurur veya yavaşlatırsa, otoriter ülkeler basitçe devam edecektir. Teknolojinin inanılmaz ekonomik ve askeri değeri ve anlamlı bir uygulama mekanizmasının olmaması göz önüne alındığında, onları durdurmaya nasıl ikna edebileceğimizi göremiyorum.

Ancak jeopolitik konusunda gerçekçi bir görüşle uyumlu, yapay zekâ geliştirmesinde hafif bir yavaşlama yolunu görüyorum. Bu yol, otokrasilerin güçlü yapay zekâya doğru ilerleyişini birkaç yıl boyunca yavaşlatmayı, yani bunu inşa etmek için ihtiyaç duydukları kaynakları,

yani çip ve yarı iletken üretim ekipmanlarını, onlardan mahrum bırakmayı içeriyor. Bu da demokratik ülkelere, güçlü yapay zekâyı daha dikkatli bir şekilde, risklerine daha fazla özen göstererek "geliştirmek" için bir tampon bölge sağlıyor; aynı zamanda otokrasileri rahatlıkla geride bırakacak kadar hızlı ilerlemelerine olanak tanıyor. Demokrasiler içindeki yapay zekâ şirketleri arasındaki yarış, endüstri standartları ve düzenlemelerin bir karışımı yoluyla, ortak bir yasal çerçeve altında ele alınabiliyor.

Anthropic, çip ihracat kontrolleri ve yapay zekânın akıllıca düzenlenmesi için baskı yaparak bu yolu çok savundu, ancak bu görünüşte sağduyulu öneriler bile Amerika Birleşik Devletleri'ndeki (bunların en önemli olduğu ülke) politika yapıcılar tarafından büyük ölçüde reddedildi. Yapay zekâ ile o kadar çok para kazanılıyor ki -kelimenin tam anlamıyla yılda trilyonlarca dolar- en basit önlemler bile yapay zekânın doğasında var olan siyasi ekonomiyi aşmakta zorlanıyor. Tuzak şu: Yapay zekâ o kadar güçlü, o kadar göz kamaştırıcı bir ödül ki, insan uygarlığının ona herhangi bir kısıtlama getirmesi çok zor.

Sagan'ın Contact'ta yaptığı gibi, bu aynı hikayenin binlerce dünyada yaşandığını hayal edebiliyorum. Bir tür bilinç kazanır, alet kullanmayı öğrenir, teknolojinin üstel yükselişine başlar, sanayileşme ve nükleer silah krizleriyle yüzleşir ve eğer bunlardan sağ çıkarsa, kumdan düşünebilen makineler yapmayı öğrenme gibi en zor ve son meydan okumayla karşı karşıya kalır. Bu sınavdan sağ çıkıp Sevgi Dolu Zarafet Makineleri'nde anlatılan güzel toplumu inşa edip edemeyeceğimiz veya köleliğe ve yıkıma boyun eğip eğmeyeceğimiz, türümüzün karakterine ve kararlılığına, ruhumuza ve özümüze bağlı olacaktır.

Birçok engele rağmen, insanlığın bu sınavı geçmek için kendi içinde güce sahip olduğuna inanıyorum. Kariyerlerini yapay zeka modellerini anlamamıza ve yönlendirmemize, bu modellerin karakterini ve yapısını şekillendirmemize yardımcı olmaya adamış binlerce araştırmacıdan cesaret ve ilham alıyorum. Bu çabaların zamanla meyve vermesi ve önemli sonuçlar doğurması için artık iyi bir şans olduğuna inanıyorum. En azından bazı şirketlerin, modellerinin biyoterörizm tehdidine katkıda bulunmasını engellemek için anlamlı ticari maliyetler ödeyeceklerini belirtmeleri beni cesaretlendiriyor. Birkaç cesur insanın, hakim siyasi rüzgarlara karşı koyarak yapay zekâ sistemlerine yönelik ilk mantıklı koruma önlemlerini içeren yasaları geçirmesinden cesaret alıyorum. Kamuoyunun yapay zekânın riskler taşıdığını anlamasından ve bu risklerin ele alınmasını istemesinden cesaret alıyorum. Dünyanın dört bir yanındaki yılmaz özgürlük ruhundan ve nerede olursa olsun tiranlığa direnme kararlılığından cesaret alıyorum.

Ancak başarılı olmak istiyorsak çabalarımızı artırmamız gerekecek. İlk adım, teknolojiye en yakın olanların insanlığın içinde bulunduğu durum hakkında gerçeği söylemesidir; bunu her zaman yapmaya çalıştım; bu makaleyle bunu daha açık ve daha acil bir şekilde yapıyorum. Bir sonraki adım, dünyanın düşünürlerini, politika yapıcılarını, şirketlerini ve vatandaşlarını bu konunun aciliyetine ve en büyük önemine ikna etmek olacaktır; yani, her gün haberlerde yer alan binlerce diğer konuya kıyasla bu konuya düşünce ve siyasi sermaye harcamaya değer olduğunu anlatmaktır. O zaman cesaretin zamanı gelecek; yeterli sayıda insan, mevcut eğilimlere karşı çıkıp, ekonomik çıkarlarına ve kişisel güvenliklerine yönelik tehditlere rağmen ilkelere bağlı kalacak.

Önümüzdeki yıllar inanılmaz derecede zor olacak, bizden verebileceğimizi düşündüğümüzden daha fazlasını isteyecek. Ancak bir araştırmacı, lider ve vatandaş olarak geçirdiğim süre boyunca, kazanabileceğimize inanmamı sağlayacak kadar cesaret ve asalet gördüm; en karanlık koşullar altında bile, insanlığın, son anda bile olsa, galip gelmek için gereken gücü ve bilgeliği toplamanın bir yolunu bulduğuna inanıyorum. Kaybedecek zamanımız yok.

Erik Brynjolfsson ve Ben Buc'a teşekkür etmek istiyorum. Hanan, Mariano-Florentino Cuéllar, Allan Dafoe, Kevin Esvelt, Nick Beckstead, Richard Fontaine, Jim McClave ve Anthropic'in birçok çalışanına bu makalenin taslakları üzerindeki yararlı yorumları için teşekkür ederim.

Dipnotlar

1 Bu, Sevgi Dolu Zarafet Makineleri'nde belirttiğim bir noktayla simetriktir; orada yapay zekanın avantajlarının bir kurtuluş kehaneti olarak düşünülmemesi gerektiğini ve somut ve gerçekçi olmanın ve büyüklenmecilikten kaçınmanın önemli olduğunu söyleyerek başlamıştım.

Sonuç olarak, kurtuluş kehanetleri ve kıyamet kehanetleri, temelde aynı nedenlerden dolayı, gerçek dünyayla yüzleşmek için yararsızdır.↩

2 Anthropic'in amacı, bu tür değişiklikler boyunca tutarlı kalmaktır. Yapay zeka riskleri hakkında konuşmak siyasi olarak popüler olduğunda, Anthropic bu risklere karşı ihtiyatlı ve kanıta dayalı bir yaklaşımı ihtiyatlı bir şekilde savundu. Yapay zekâ riskleri hakkında konuşmak artık siyasi olarak popüler olmadığı için, Anthropic bu risklere karşı ihtiyatlı ve kanıta dayalı bir yaklaşımı savunmaya devam ediyor.↩

3 Zamanla, yapay zekânın gidişatına ve her alanda insan yeteneğini aşma olasılığına olan güvenim giderek arttı, ancak yine de bazı belirsizlikler mevcut.↩

4 Çip ihracat kontrolleri bunun harika bir örneği. Basitler ve çoğunlukla işe yarıyor gibi görünüyorlar.↩

5 Ve elbette, bu tür kanıtların aranması entelektüel olarak dürüst olmalı, böylece tehlikenin olmadığına dair kanıtlar da ortaya çıkabilir. Model kartları ve diğer açıklamalar yoluyla şeffaflık, bu tür entelektüel açıdan dürüst bir girişimin bir örneğidir.↩

6 Gerçekten de, 2024'te "Sevgi Dolu Zarafet Makineleri"ni yazdıktan bu yana, yapay zeka sistemleri insanların birkaç saatini alan görevleri yapabilir hale geldi; METR yakın zamanda Opus 4.5'in %50 güvenilirlikle yaklaşık dört insan saatlik işi yapabileceğini değerlendirdi.↩

7 Ve açıkça söylemek gerekirse, güçlü yapay zeka teknik anlamda sadece 1-2 yıl uzakta olsa bile, hem olumlu hem de olumsuz birçok toplumsal sonucu birkaç yıl daha sürebilir. Bu nedenle, yapay zekanın 1-5 yıl içinde giriş seviyesi beyaz yakalı işlerin %50'sini alt üst edeceğini düşünürken, aynı zamanda sadece 1-2 yıl içinde herkesten daha yetenekli bir yapay zekaya sahip olabileceğimizi de düşünebiliyorum.↩

8 Kamuoyunun (politika yapıcılara kıyasla) yapay zeka riskleriyle çok endişeli göründüğünü de eklemekte fayda var. Bence odak noktalarının bir kısmı doğru (örneğin yapay zeka kaynaklı iş kayıpları), bir kısmı ise yanlış yönlendirilmiş (örneğin yapay zekanın su kullanımına ilişkin endişeler, ki bu önemli değil). Bu tepki, riskleri ele alma konusunda bir fikir birliğinin mümkün olduğuna dair bana umut veriyor, ancak şimdiye kadar bu, etkili veya iyi hedeflenmiş politika değişikliklerinden bahsetmiyorum bile, politika değişikliklerine henüz dönüşmedi.↩

9 Elbette, fiziksel dünyada istediklerini yapmaları için çok sayıda insanı manipüle edebilirler (veya basitçe para ödeyebilirler).↩

10 Bunun bir safsata olduğunu düşünmüyorum: Örneğin, Yann LeCun'un bu görüşü savunduğunu anlıyorum.↩

11 Örneğin, Claude 4 sistem kartının 5.5.2 bölümüne (s. 63–66) bakın.↩

12 Basit modelde, burada tartışmayacağım bir dizi başka varsayım da var. Genel olarak, bu durum bizi yanlış hizalanmış güç arayışının basit öyküsü konusunda daha az endişelendirmeli, ancak öngöremediğimiz olası öngörülemeyen davranışlar konusunda daha fazla endişelendirmelidir.↩

13 Ender's Game, yapay zekâ yerine insanları içeren bir versiyonunu anlatıyor.↩

14 Örneğin, modellere çeşitli kötü şeyler yapmamaları ve aynı zamanda insanlara itaat etmeleri söylenebilir, ancak daha sonra birçok insanın tam olarak bu kötü şeyleri yaptığını gözlemleyebilirler! Bu çelişkinin nasıl çözüleceği açık değil (ve iyi tasarlanmış bir anayasa, modeli bu çelişkileri zarif bir şekilde ele almaya teşvik etmelidir), ancak bu tür bir ikilem, test sırasında yapay zekâ modellerini içine koyduğumuz sözde "yapay" durumlardan çok farklı değil.↩

15 Bu arada, anayasanın doğal dil belgesi olmasının bir sonucu da, dünyanın okuyabileceği ve bu da herkes tarafından eleştirilebileceği ve diğer şirketlerin benzer belgeleriyle karşılaştırılabileceği anlamına gelmesidir. Şirketleri bu belgeleri yayınlamaya teşvik etmekle kalmayıp, aynı zamanda iyi olmaya da teşvik eden bir "en iyiye doğru yarış" yaratmak değerli olurdu.↩

16 Anayasal Yapay Zeka'dan gelen karakter tabanlı yaklaşımı, yorumlanabilirlik ve uyum bilimi sonuçlarına bağlayan derin bir birleştirici ilke hakkında bir hipotez bile var. Hipoteze göre, Claude'u yönlendiren temel mekanizmalar başlangıçta, bir romandaki karakterlerin ne söyleyeceğini tahmin etmek gibi, ön eğitimde karakterleri simüle etme yolları olarak ortaya çıktı. Bu, anayasayı düşünmenin yararlı bir yolunun, modelin tutarlı bir kişiliği somutlaştırmak için kullandığı bir karakter tanımı gibi olduğunu öne sürer. Ayrıca yukarıda bahsettiğim "Ben kötü bir insan olmalıyım" sonuçlarını açıklamamıza yardımcı olur (çünkü model tutarlı bir karaktermiş gibi davranmaya çalışıyor - bu durumda kötü bir karakter) ve yorumlanabilirlik yöntemlerinin modeller içindeki "psikolojik özellikleri" keşfedebileceğini öne sürer. Araştırmacılarımız bu hipotezi test etmenin yolları üzerinde çalışıyorlar.

17 Açık olmak gerekirse, izleme gizliliği koruyacak şekilde yapılır.

18 Sorumlu Ölçeklendirme Politikamızla esasen gönüllü olarak dayatılan kurallarla yaptığımız kendi deneylerimizde bile, önceden önemli görünen ancak geriye dönüp bakıldığında saçma olduğu ortaya çıkan çizgiler çizerek çok katı olmanın çok kolay olduğunu tekrar tekrar gördük. Bir teknoloji hızla ilerlerken yanlış şeyler hakkında kurallar koymak çok kolaydır.

19 SB 53 ve RAISE, yıllık geliri 500 milyon doların altında olan şirketler için hiç geçerli değildir. Sadece Anthropic gibi daha büyük, daha köklü şirketler için geçerlidir.
20 Joy'un makalesini ilk olarak 25 yıl önce, yazıldığında okudum ve üzerimde derin bir etkisi oldu. O zaman da şimdi de, bunu çok karamsar buluyorum—Joy'un önerdiği gibi, teknolojinin tüm alanlarından geniş çaplı "vazgeçmenin" çözüm olduğunu düşünmüyorum—ancak ortaya koyduğu sorunlar şaşırtıcı derecede öngörülüydü ve Joy ayrıca hayran olduğum derin bir şefkat ve insanlık duygusuyla yazıyor.↩
21 Devlet aktörleri konusunda, şimdi ve gelecekte endişelenmemiz gerekiyor ve bunu bir sonraki bölümde ele alıyorum.↩
22 Birçok teröristin en azından nispeten iyi eğitimli olduğuna dair kanıtlar var; bu, burada yetenek ve motivasyon arasında negatif bir korelasyon olduğunu savunduğum şeyle çelişiyor gibi görünebilir. Ancak bence gerçekte bunlar uyumlu gözlemlerdir: başarılı bir saldırı için yetenek eşiği yüksekse, o zaman neredeyse tanım gereği, şu anda başarılı olanların, yetenek ve motivasyon arasında negatif bir korelasyon olsa bile, yüksek yeteneğe sahip olmaları gerekir. Ancak yetenek üzerindeki sınırlamaların kaldırıldığı bir dünyada (örneğin, gelecekteki LLM'lerle), öldürme motivasyonuna sahip ancak yeteneği düşük olan önemli bir nüfusun bunu yapmaya başlayacağını tahmin ediyorum - tıpkı çok fazla yetenek gerektirmeyen suçlarda (okul saldırıları gibi) gördüğümüz gibi.↩
23 Ancak Aum Shinrikyo denedi. Aum Shinrikyo'nun lideri Seiichi Endo, Kyoto Üniversitesi'nden viroloji eğitimi almıştı ve hem şarbon hem de ebola üretmeye çalıştı. Ancak 1995 itibariyle, o bile bunu başarmak için

yeterli uzmanlığa ve kaynaklara sahip değildi. Çıta artık önemli ölçüde daha düşük ve LLM'ler bunu daha da düşürebilir.↩

24 Seri katillerle ilgili tuhaf bir olgu, seçtikleri cinayet tarzının neredeyse grotesk bir moda gibi işlemesidir. 1970'lerde ve 1980'lerde seri katiller çok yaygındı ve yeni seri katiller genellikle daha yerleşik veya ünlü seri katillerin davranışlarını kopyalıyordu. 1990'larda ve 2000'lerde toplu silahlı saldırılar daha yaygın hale gelirken, seri katiller daha az yaygın hale geldi. Bu davranış kalıplarını tetikleyen teknolojik bir değişim yok; sadece şiddet uygulayan katillerin birbirlerinin davranışlarını kopyaladığı ve kopyalanacak "popüler" şeyin değiştiği görülüyor.

25 Bazen sıradan hapishaneden kaçanlar, modelin bir virüsün genom dizisi gibi belirli bir bilgi parçasını üretmesini sağladıklarında bu sınıflandırıcıları tehlikeye attıklarına inanırlar. Ancak daha önce açıkladığım gibi, endişelendiğimiz tehdit modeli, biyolojik silah üretim sürecindeki belirli belirsiz adımlar hakkında haftalar veya aylarca süren adım adım, etkileşimli tavsiyeler içeriyor ve sınıflandırıcılarımız da buna karşı savunmayı amaçlıyor. (Araştırmamızı genellikle “evrensel” hapisten çıkış yolları aramak olarak tanımlıyoruz; bunlar sadece belirli veya dar bir bağlamda işe yaramayan, modelin davranışını geniş ölçüde açan yollar.)↩

26 Sınıflandırıcılarımızı daha verimli hale getirmek için yatırım yapmaya devam edeceğiz ve şirketlerin bu gibi ilerlemeleri birbirleriyle paylaşmaları mantıklı olabilir.↩

27 Açıkçası, şirketlerin engelledikleri biyolojik silah üretimindeki belirli adımlar hakkında teknik ayrıntıları açıklamak zorunda kalmaları gerektiğini düşünmüyorum ve şimdiye kadar kabul edilen şeffaflık mevzuatı (SB 53 ve RAISE) bu konuyu ele alıyor.↩

28 Bir diğer ilgili fikir ise, hükümetin acil bir durumda bu ekipman için önceden kararlaştırılmış bir fiyat ödeme sözü vererek, biyolojik bir saldırıya yanıt vermek için gerekli olan KKD, solunum cihazları ve diğer temel ekipmanların stoklanmasını teşvik ettiği “dayanıklılık piyasaları”dır. Bu, tedarikçileri, hükümetin tazminat ödemeden el koyacağından korkmadan bu tür ekipmanları stoklamaya teşvik eder.↩

29 Neden büyük aktörlerin iktidarı ele geçirmesinden, küçük aktörlerin ise yıkıma neden olmasından daha çok endişeleniyorum? Çünkü dinamikler farklı. İktidarı ele geçirmek, bir aktörün diğer herkesi alt etmek için yeterli gücü toplayıp toplayamayacağıyla ilgilidir; bu nedenle en güçlü aktörler ve/veya yapay zekaya en yakın olanlar hakkında endişelenmeliyiz. Buna karşılık, yıkım, savunması yapmaktan çok daha zorsa, az güce sahip olanlar tarafından da gerçekleştirilebilir. O zaman bu, muhtemelen daha küçük aktörler olan en çok sayıda tehdide karşı savunma oyunudur.↩

30 Bu, saldırı ve savunmanın siber saldırılarda biyolojik silahlara göre daha dengeli olabileceği yönündeki görüşümle çelişkili gibi görünebilir, ancak buradaki endişem, bir ülkenin yapay zekası dünyanın en güçlüsü ise, diğerlerinin de aynı şekilde güçlü olabileceğidir. Teknolojinin kendisinde içsel bir saldırı-savunma dengesi olsa bile, kullanıcılar kendilerini savunamayacaklardır.↩

31 Örneğin, Amerika Birleşik Devletleri'nde bu, dördüncü maddeyi ve Posse Comitatus Yasasını içerir.↩

32 Ayrıca, açıkça belirtmek gerekirse, özellikle demokrasilerdeki şirketler tarafından kontrol ediliyorsa, farklı yönetim yapılarına sahip ülkelerde büyük veri merkezleri inşa etmenin bazı gerekçeleri vardır. Bu tür inşaatlar, prensipte demokrasilerin daha büyük tehdit olan Çin Komünist Partisi ile daha iyi rekabet etmelerine

yardımcı olabilir. Ayrıca, bu tür veri merkezlerinin çok büyük olmadıkları sürece fazla risk oluşturmadığını düşünüyorum. Ancak genel olarak, kurumsal güvencelerin ve hukukun üstünlüğü korumalarının daha az yerleşik olduğu ülkelerde çok büyük veri merkezleri yerleştirirken ihtiyatlı olunması gerektiğini düşünüyorum.↩

33 Bu, elbette, nükleer caydırıcılığın güvenliğinin iyileştirilmesi ve güçlü yapay zekaya karşı daha sağlam hale getirilmesi için de bir gerekçedir ve nükleer silahlı demokrasiler bunu yapmalıdır. Ancak güçlü bir yapay zekanın neler yapabileceğini veya ona karşı hangi savunmaların işe yarayacağını bilmiyoruz, bu nedenle bu önlemlerin sorunu mutlaka çözeceğini varsaymamalıyız.↩

34 Nükleer caydırıcılık etkili kalsa bile, saldıran bir ülkenin blöfümüzü ortaya çıkarmaya karar verebileceği riski de var; insansız hava aracı sürüsünün bizi fethetme riski önemli olsa bile, nükleer silahları bir insansız hava aracı sürüsüne karşı savunmak için kullanmaya istekli olup olmayacağımız belirsiz. İnsansız hava aracı sürüleri, nükleer saldırılardan daha az şiddetli ancak konvansiyonel saldırılardan daha şiddetli yeni bir şey olabilir. Alternatif olarak, yapay zeka çağında nükleer caydırıcılığın etkinliğine ilişkin farklı değerlendirmeler, nükleer çatışmanın oyun teorisini istikrarsızlaştırıcı bir şekilde değiştirebilir.↩

35 Açık olmak gerekirse, güçlü yapay zekaya ulaşma süresi önemli ölçüde daha uzun olsa bile, Çin'e çip satmamanın doğru strateji olduğuna inanıyorum. Çinlileri Amerikan çiplerine "bağımlı" hale getiremeyiz; yerli çip endüstrilerini bir şekilde geliştirmeye kararlılar. Bunu başarmaları yıllar alacak ve biz onlara çip satarak sadece bu süre zarfında onlara büyük bir destek sağlıyoruz.↩

36 Açık olmak gerekirse, bugün Ukrayna ve Tayvan'da kullanılanların çoğu tamamen otonom silahlar değil. Bunlar geliyor, ancak bugün burada değiller.↩
37 En yeni modelimiz Claude Opus 4.5 için model kartımız, Opus'un Anthropic'te sık sık verilen bir performans mühendisliği mülakatında şirketin tarihindeki herhangi bir mülakatçıdan daha iyi performans gösterdiğini ortaya koyuyor.↩
38 "Tüm kodu yazmak" ve "bir yazılım mühendisinin görevini baştan sona yapmak" çok farklı şeylerdir, çünkü yazılım mühendisleri sadece kod yazmaktan çok daha fazlasını yaparlar; test etme, ortamlarla, dosyalarla ve kurulumla ilgilenme, bulut bilişim dağıtımlarını yönetme, ürünler üzerinde yineleme yapma ve daha fazlasını içerir.↩
39 Bilgisayarlar bir anlamda geneldir, ancak birkaç alanda (aritmetik gibi) insanlardan çok daha üstün olsalar bile, insan bilişsel yeteneklerinin büyük çoğunluğuna kendi başlarına sahip olamazlar. Elbette, yapay zeka gibi bilgisayarlar üzerine kurulu şeyler (programlar) artık çok çeşitli bilişsel yeteneklere sahip; bu makale de tam olarak bununla ilgili.↩
40 Açık olmak gerekirse, yapay zeka modelleri insanlarla tam olarak aynı güçlü ve zayıf yönler profiline sahip değil. Ancak her boyutta oldukça düzgün bir şekilde ilerliyorlar, bu nedenle sivri veya düzensiz bir profile sahip olmanın nihayetinde bir önemi olmayabilir.↩
41 Bununla birlikte, ekonomistler arasında bu fikir hakkında tartışmalar var.↩
42 Kişisel servet bir "stok" iken, GSYİH bir "akış"tır, bu nedenle bu, Rockefeller'ın Amerika Birleşik Devletleri'ndeki ekonomik değerin %2'sine sahip olduğu iddiası değildir. Ancak bir ulusun toplam servetini ölçmek GSYİH'den daha zordur ve insanların bireysel gelirleri

yıldan yıla çok değişir, bu nedenle aynı birimlerde bir oran yapmak zordur. En büyük kişisel servetin GSYİH'ye oranı, elma ile elmayı karşılaştırmasa da, aşırı servet yoğunlaşması için son derece makul bir ölçüttür.↩

43 Ekonomideki toplam emek değeri yılda 60 trilyon dolardır, bu nedenle yılda 3 trilyon dolar bunun %5'ine karşılık gelir. Bu miktar, insan maliyetinin %20'si karşılığında emek sağlayan ve %25 pazar payına sahip bir şirket tarafından kazanılabilir, hatta emek talebi artmasa bile (ki düşük maliyet nedeniyle neredeyse kesinlikle artacaktır).↩

44 Açık olmak gerekirse, gerçek yapay zeka verimliliğinin henüz ABD ekonomik büyümesinin önemli bir bölümünden sorumlu olduğunu düşünmüyorum. Aksine, veri merkezi harcamalarının, piyasanın gelecekteki yapay zeka odaklı ekonomik büyümeyi beklemesinden ve buna göre yatırım yapmasından kaynaklanan öngörücü yatırımdan kaynaklanan büyümeyi temsil ettiğini düşünüyorum.↩

45 Yönetimle aynı fikirde olduğumuzda bunu söylüyoruz ve karşılıklı olarak desteklenen politikaların dünya için gerçekten iyi olduğu noktalarda mutabakat arıyoruz. Biz, herhangi bir siyasi partinin destekçisi veya karşıtı olmaktan ziyade dürüst arabulucular olmayı hedefliyoruz.

46 Birkaç yıldan daha uzun bir sürenin mümkün olduğunu düşünmüyorum: daha uzun zaman dilimlerinde kendi çiplerini üretecekler.